IMPROVING OCCUPATIONAL EXPOSURE ESTIMATES THROUGH MULTIPLE IMPUTATION TO REDUCE EXPOSURE MISCLASSIFICATION

by

Pamela Jean Dopart

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

September 2015

ABSTRACT

Exposure misclassification is present in nearly every occupational and environmental epidemiology study and, if left unaddressed, can bias risk estimates and exposure-response relationships of interest. The potential for exposure misclassification may be increased when the measurement data used to define exposure are limited or incomplete. Multiple imputation is a technique for addressing missing data that has many advantages, including retention of all available data and preservation of the population variability. The objective of this research was to reduce the potential for misclassification due to missing data by improving exposure estimates through a multiple imputation approach. This objective was explored through three separate research aims: to understand and characterize common missing exposure data patterns in occupational cohorts; to comment on the observed influence missing data patterns have on the ability to accurately estimate exposures of a work population; and to test the performance of a multiple imputation approach in characterizing exposures under predetermined missing data scenarios.

Using a comprehensive and complete dataset of radiation exposures of naval shipyard workers, missing data were artificially and purposefully generated under several hypothesized occupational sampling scenarios. The analyses were divided into three separate chapters based on the populations for which exposures were estimated. Population-level exposures within a given shipyard were examined in Chapter 3. The homogeneity of exposures within a purported similar exposure group (SEG) was explored in Chapter 4. Finally, in Chapter 5, the ability to develop exposure levels for a specific shipyard of interest by using surrogate data from separate shipyards was investigated.

Overall, the multiple imputation approach performed well in estimating exposure levels for the population of interest, even when the percentage of missing data was very high (greater than 95% missing). By simulating various plausible sampling plans, several general remarks can be made. Exposure levels that decrease over time is a commonly observed pattern in occupational cohorts; however, this occurrence can impact the feasibility of combining measurements from multiple time periods when data are limited. Biased sampling plans in which workers are intentionally measured (or not measured) based on the perceived exposure levels of their job titles are also a common practice but were not shown to significantly impact the ability to estimate exposures. Heterogeneity of exposures within an SEG can be significant; understanding the strongest determinants of exposure can assist in developing improved exposure groupings. Finally, when estimating exposures for a specific facility by combining data from multiple locations, the impact on exposure estimates due to the differences between facilities can be attenuated by combining data from as many different facilities and over as long a time period as available.

The results from these analyses allow for a better understanding of some of the most characteristic missing data patterns observed in occupational cohorts and of the impact these missing data have on the ability to accurately assess exposure. In addition, multiple imputation proved to be a viable tool for addressing missing exposures within occupational datasets. This research highlights the importance of exploring the potential differences between populations with available exposure data and those without and of accounting for those differences when generating exposure estimates.

ACKNOWLEDGEMENTS

I would like to start by thanking Peter Lees, my advisor and mentor over these past five years. Peter taught me so much more about the field of occupational exposure assessment than I could have ever imagined and I look forward to paying that gift back through my own future contributions to the field. Thank you also for always having an open door and being willing to talk about anything, big or small. “The data ARE” is forever burned inside my brain.

I would also like to thank the other members of my final defense committee: Frank Curriero, who has been involved since the beginning of my research and has always been willing to help, Kirsten Koehler, and Marie Diener-West. I am also grateful to Ana Navas-Acien, who was on my thesis committee and offered insightful comments during our meetings every six months, and to Elizabeth Colantuoni, who provided helpful guidance on some of the biostatistics work.

I want to thank Elsbeth Chee for her help in handling and interpreting the radiation datasets and for making sure our office socialized! I would also like to thank Gene Matanoski and the rest of the Shipyard Study Team for all of your help: Greg Surplus, Wendy Pichardo, Cilicia Lawson, Charlie Klassen, and Linda Schwartz. Thank you SO much to the EHS staff for all your help in getting me to this point and for everything you do for our department, especially Courtney Mish, Nina Kulacki, Ruth Quinn, and Patty Poole. I am also so glad that EHSSO is alive and well and something I got to be a part of for three years.

Thank you to all of my EHS friends for keeping me sane, letting me vent, and providing me with good memories of the city of Baltimore. An extra thanks goes to my officemates for all of

the above, especially Sut Soneja, Stacy Woods, and Ben Davis (and Jesse Negherbon, even though he was never there). I am also grateful for my IH girls for cheering each other on during major life events and for letting me sometimes forget that I was the only one still in school: Rachel Seymour, Michelle Coutu, and Debbie Greene.

I also want to thank two of my best friends in the world, Michelle Ginfrida and Melissa Bondar, who have in many ways been on this journey with me. I may have missed out on some adventures over the past few years but I never felt left out. And Melissa, thank you for giving us the opportunity to jump out of an airplane!

I am extremely grateful to Baltimore for introducing me to Tyler Morgan-Wall. There is no one else I would rather have gone on this journey with and the fact that we went through it together makes it that much sweeter. Thank you not only for your emotional support and love but also for your programming skills and technical advice, which (no exaggeration) got my project going. I cannot wait to marry you.

Finally, thank you to my parents, Alan and Cathy Dopart, and my sister Eileen Dopart. I could not have done this without the love and support I've constantly received from you, no matter what. My parents have been nothing but supportive, always enthusiastic about the bigger picture, even when I sometimes couldn't see it. Thank you for the E-Z Pass too, so I could come home for respite when needed! And thank you Eileen for all the hilarious phone calls; somehow you seemed to always know when I needed a laugh. And finally, thank you to Kayla Dopart, my four-legged family member, for always being right by my side.

THESIS READERS

Thesis Advisor

Peter SJ Lees, Professor, Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Thesis Committee

Frank Curriero, Associate Professor, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Marie Diener-West, Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Kirsten Koehler, Assistant Professor, Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2: Methods

Chapter 3: Estimating Population-level Exposures

Chapter 4: Characterizing the Exposure Profile of an SEG

Chapter 5: Developing Exposure Estimates Using Surrogate Data

CHAPTER 1: INTRODUCTION

1.1. Background

1.1.1. Exposure Assessment

In trying to elucidate a relationship between an exposure of concern and a particular outcome, a human-health risk assessment is often carried out. Such an assessment, in general, involves the evaluation of available scientific information on the hazardous properties of the agent of concern and on the extent of human exposure to this agent. The paradigm is typically considered in five formal steps: problem scoping, hazard identification, dose-response assessment, exposure assessment, and risk characterization. Integration of information from these steps allows for development of a risk estimate reflecting the likelihood that the outcome under study will occur in the exposed population (NRC, 1983; NRC, 2009).

While meaningful risk estimates are dependent on each step of the paradigm, considerable weight should be given to the importance of a thorough exposure assessment. Exposure assessments have applications beyond the risk assessment paradigm as well – including compliance determinations, health complaints, and epidemiologic studies – all of which are greatly affected by the quality of the assessment (Stewart & Stenzel, 2000). Formally, exposure assessment is defined as the evaluation of the intensity, frequency, and duration of contact between a hazard and the external boundaries of the human body (Fed. Reg., 1992). A 2000 publication by Stewart and Stenzel suggested that exposure assessment can be thought to have five components:

collection of data, identification of the hazard, selection of exposure metrics, definition of exposure groups, and estimation of the exposures (Steward & Stenzel, 2000). The quality of the overall exposure assessment is dependent upon each of these steps; any sources of error or uncertainty associated with a particular component will be reflected in the final exposure estimates. If not properly identified and accounted for, these sources can lead to inaccurate estimations, which in turn can result in exposure misclassification of the study population.

1.1.2. Exposure Misclassification

Misclassification is defined as the "erroneous classification of an individual, a value, or an attribute into a category other than that to which it should be assigned" (Last, 2001). Exposure misclassification can bias the interpretation of an exposure-outcome association, as risk estimates are determined using exposure classifications that may be incorrect. Previous research has attempted to describe the impact of exposure misclassification (Armstrong, 1998; Blair et al. 2007; Copeland et al. 1977; Jurek et al. 2005; Jurek et al. 2008; Weinkam et al. 1991). The magnitude and direction of the effect of misclassification on estimates of risk were shown to vary by the degree of misclassification, the presence of other biases, and the prevalence of the exposure; however, these studies concluded that even relatively small errors could have sizable effects on the risk estimates. Given the potential public health policy implications of a completed risk assessment, the importance of developing quality risk estimates cannot be understated. Exposure misclassification can also have significant consequences beyond the scope of a formal risk assessment. When assessing the exposures of an occupational

cohort, for instance, misclassified exposures can have an unintended impact on decisions regarding worker protection policies within a given company or across an industry.

1.1.3. Missing Data

The potential for exposure misclassification may be increased when the measurement data used to define exposure are limited or incomplete. Missing data are a common and often unavoidable issue; however, the fact that they can hinder data analysis procedures necessitates that they be addressed. Deciding how to approach missing values in a dataset first requires an understanding of the mechanisms by which data become missing (Schafer & Graham, 2002).

Missing data patterns are broadly classified into three categories, or missing data mechanisms (Rubin, 1976; van Buuren, 2012). If the probability of being missing is the same for all cases, and the data are missing independently of both the observed and unobserved data, then the data are said to be missing completely at random (MCAR) and thus the reasons for the missing data are unrelated to the data itself. While the most convenient mechanism for analysis purposes, MCAR is often an unrealistic scenario. If the probability of being missing is based on the observed values of the other variables, but still independent of the unobserved data, then the data are considered to be missing at random (MAR). Many modern missing data methods, including multiple imputation, start from this assumption. Finally, if the data are missing because of reasons related to the values of the unobserved data, then the data are considered to be missing not at random

(MNAR) (Rubin, 1976; van Buuren, 2012). Unlike under MCAR or MAR missingness, analyses under MNAR missingness yield biased parameter estimates (Graham, 2009).

The impact of the missing data on the results of a statistical analysis is determined by both the true missing data mechanism and by the assumptions made by the analyst when choosing a course of action for addressing the missing values. In practice, however, it can be quite difficult, if not impossible, to identify the correct missing data mechanism for a dataset, particularly when attempting to distinguish between MAR and MNAR (Graham, 2009).

1.1.4. Techniques for Addressing Missing Data

There are a number of strategies for addressing missing data (Pigott, 2001). Each method requires assumptions about the nature of the data and the reasons for the missing observations. When a particular approach is employed without careful consideration of the assumptions required of that method, there exists a risk of obtaining biased or misleading results. It can be difficult to choose the best appropriate method for a given dataset; there are indeed scenarios in which each of the approaches described below are appropriate. However, some missing data methods have emerged as preferred strategies given the advantages they have over other techniques; this includes the method selected for this dissertation, multiple imputation (Greenland & Finkle, 1995).

Deletion methods, such as complete-case analysis (CCA), analyze only those cases with complete information and drop those with missing data. This approach is commonly

applied due to its simplicity of use; however, a major disadvantage of CCA is the loss of data and thus a decrease in the overall sample size. This can lead to reductions in statistical power and, depending on the mechanism of missingness and whether the complete cases are truly a random sample, to biased results (Demissie et al. 2003; Enders, 2011). Another common strategy is to substitute one plausible value, such as the mean of the observed cases, for all missing observations. The main advantage of this approach over a deletion method is that is allows for retention of all the data; however, by substituting a single value for the missing data, there is an underestimation of the true variance (Pigott, 2001).

Model-based methods, such as maximum likelihood and multiple imputation, are more sophisticated and computationally advanced and offer advantages over the alternative methods. By incorporating information on the partially observed cases, these techniques can be considered unbiased and statistically more powerful analyses as compared to the ones described above, especially when the remaining variables are good predictors of the variable(s) with missing data (Raghunathan 2004; Sterne et al. 2009). Including Furthermore, the missing data mechanism assumed when using these techniques is less restrictive than would be required for a technique such as CCA as data are assumed to be MAR rather than MCAR (Pigott, 2001). Since this research tests the performance of a multiple imputation method, a more detailed description of this technique is provided in Chapter 2.

1.1.5. Missing Occupational Exposure Data

The previous discussions on exposure misclassification and missing data apply to a wide range of exposure data. This dissertation focuses on the workplace exposures of an occupational cohort, which, in addition to the above concerns, has its own set of unique challenges. Both industrial hygienists and occupational epidemiologists rely on available data to characterize as accurately as possible the exposure profiles of a work population (Stewart and Stenzel, 1999). However, the manner in which the data are used and interpreted can vary significantly based on the questions the data handler wishes to answer. Industrial hygienists most often monitor the work population to ensure compliance with regulatory occupational exposure limits (OELs) and thus design their sampling plans accordingly (Harris, 1995). For example, the hygienist may choose to purposefully oversample from those workers identified as having the highest exposures, particularly when limited in the total number of samples that can be collected (Checkoway et al. 2004). Doing so gives the hygienist the ability to confidently determine that all workers are exposed below the required exposure limits. Epidemiologists and other researchers, on the other hand, strive to optimize exposure estimates with the aim of detecting a possible risk and/or characterizing an exposure-response association. Risks associated with occupational exposures can often be small; to detect a risk that is truly there, the exposure assessment need to be quite refined (Nieuwenhuijsen, 2003).

Researchers investigating occupational exposures are often dependent on the sampling schemes originally developed by industrial hygienists for the purposes of compliance

determinations. Thus, the questions being asked of the data may be quite different from the ones considered when the sampling plan was designed. A comprehensive industrial hygiene sampling plan ideally includes enough information to assess exposure for an occupational epidemiology study (Checkoway et al. 1987; Harris, 1995; Stewart and Stenzel, 1999). However, before analyzing exposure data for their own purposes, researchers should be mindful of the original purpose of the collected data. This includes any missing data patterns that might be observed. The amount of data missing – and the characteristics of the workers with missing exposures – may vary based on the hygienists' initial expectations of the exposure profile and/or the overall goal of their sampling plan.

When faced with addressing missing exposure data in an occupational cohort, the researcher has many options, including the ones described above. However, the performance of these approaches can be altered by missing data patterns, especially ones of which the researcher may be unaware. Such analytical techniques require an understanding of how the data are missing and how the observed data will influence any generated exposure estimates. For an approach such as multiple imputation to be useful in addressing missing occupational exposure data, common missing data patterns in occupational cohorts should first be understood. The performance of the statistical approach can then be evaluated using a dataset containing artificially missing data that are generated in ways that reflect the observed common missing data patterns. If missing data can be properly addressed, and accurate exposure estimates can be developed, then the potential for exposure misclassification will be reduced.

1.2. Research Aims

The research in this dissertation aimed to reduce the potential for exposure misclassification by improving exposure estimates through a multiple imputation approach. Using a comprehensive and complete dataset containing radiation exposures of naval shipyard workers, in which missing data were artificially and purposefully generated, the performance of a multiple imputation technique for estimating missing exposure values under several hypothesized sampling scenarios was examined. The overall aims of this research included: to understand and characterize common missing exposure data patterns in occupational cohorts; to test the performance of a multiple imputation approach in characterizing exposures under predetermined missing data scenarios; and to comment on the observed influence missing data patterns have on the ability to accurately estimate exposures of a work population.

The analyses in this dissertation are divided into three chapters based on the population for which exposures are estimated. The specific aims of Chapter 3, *Estimating Population-level Exposure*, are to understand common missing data patterns in occupational cohorts, examine the effect these patterns have on the ability to accurately characterize population-level exposures, and to test the performance of a multiple imputation approach in estimating such population exposures. In Chapter 4, *Characterizing the Exposure Profile of an SEG*, the specific aims are to examine how are similar exposure groups (SEGs) are affected by various sampling plans, explore additional workplace variables that may influence the homogeneity of an SEG, and test

the performance of a multiple imputation approach in estimating SEG-level exposures. Finally, the specific aims of Chapter 5, *Developing Exposure Estimates Using Surrogate Data*, are to compare between shipyards the exposure profile of naval shipyard workers during various time periods and to test the performance of a multiple imputation approach in estimating exposure levels when surrogate exposure data are used.

The results of these analyses will allow for a better understanding of some of the most characteristic missing data patterns observed in occupational cohorts, of the impact these missing data have on the ability to accurately assess exposure, and of the potential application of multiple imputation in estimating occupational exposures.

CHAPTER 2: METHODS

2.1. Study Population

The study population used for the analyses in this dissertation was selected due to the size and completeness of the exposure records. Over one million exposure measurements collected on approximately 13,800 employees were available for the investigations described throughout Chapters 3-5. By working with a large dataset that contains no missing data, many plausible missing data patterns observed in occupational cohorts were simulated.

2.1.1. Shipyard Population

The overall shipyard population represented all employees, both nuclear and non-nuclear workers, in eight shipyards (a combination of U.S. Navy and private yards) and was established for the purposes of an epidemiology study investigating the cancer risk of low-level radiation to U.S. shipyard workers (Matanoski et al. 2008). For this dissertation, shipyard employees who were 1) considered to be radiation exposed workers and 2) employed at one of a selected three yards were assembled as the initial study population, on which the selection criteria described below were applied.

Radiation exposed workers were defined as all employees certified to work in areas that, at some time beginning with the start of nuclear ship overhauls, had the potential for exposure to radioactivity and who had an employment record and a dosimetry record in the radiation database (Matanoski et al. 2008). Ever since overhauls of nuclear powered ships began, workers in the selected shipyards have been monitored for radiation

exposure while working in the reactor areas. Overhauls began at different times within each yard during the 1957-1967 timeframe; however, similar radiation monitoring programs were implemented in each shipyard with the start of its initial overhauls. The shipyards required constant monitoring of each worker to ensure compliance with standards imposed by the U.S. Navy; thus, this radiation exposed population includes measured exposures for every individual over the entire period of overhauls (Matanoski et al. 2008, Correa). Workers were instructed to wear a radiation dosimeter whenever they entered a potential radiation exposure area, regardless of the actual resulting exposure level. For this dissertation, these exposures are reflected in a database that contains radiation exposures of approximately 13,800 radiation workers over a study period of 1975 to 2005.

2.1.2. Selection Criteria

For the purposes of this research several selection criteria were applied to the original shipyard population. Only radiation measurements that were reported as having been received at the shipyard under study were included. In order to focus the analyses on workers in the skilled trades, who would be most likely to perform repair and overhaul work on the vessels, restrictions were placed on the agency to which a worker was employed and the occupational category to which a worker's job title belonged. Only workers who were employed in the Naval Sea Systems Command agency (NV24) or the U.S. Pacific Fleet, Command in Chief (NV70; applies to NS3 only) and had an occupational category code of B, or *blue collar*, were selected. Finally, given the small percentage of female workers in the population, only male workers were included. The

same selection criteria were applied to all three shipyards, which are described in detail below. The populations used in the various analyses described throughout this dissertation have been selected from within this finalized study population. When the goal of an exercise was to estimate the overall exposure levels of the shipyard population (the "population-level" exposures), the entire study population was used. When the goal was to estimate the exposures for a specific job title or SEG (the "SEG-level" exposures), then those specific job titles were extracted from the overall study population.

2.1.3. Summary of Work Population at Shipyard #1

Following the selection criteria, the study population for Naval Shipyard #1 (NS1) consisted of 440,463 daily and 43,462 annual radiation records (the difference between the two types of records is detailed below), which characterized the exposures of 6,105 workers. As shown in Table 2.1, 71% of the 6,105 workers were white. The highest level of education for nearly two-thirds of the workers was high school graduate; approximately 25% of workers completed a terminal occupational program in a technical or skilled field or attended some college. Nearly 70% of the workers were born during or after 1950.

The NS1 radiation dataset spans from 1975 to 2005 (Table 2.2). There were 86 unique job titles on which radiation measurements were collected. Of the 440,463 daily measurements collected, 50.3% had a value of 0 mrem. The population mean exposure level was 15.4 mrem and the population median exposure level was 0.0 mrem. Of the 43,462 annual measurements, 23.2% had a value of 0 mrem. The annual population mean

exposure level was 156.7 mrem and the median exposure level was 18.0 mrem. The observed increase in exposure levels in the annual dataset is expected, as each annual record is a summation of all daily exposure measurements that were collected on a given worker in a given year. On average, approximately 10 daily measurements were collected on a given worker during a one year period.

2.1.4. Summary of Work Population at Shipyard #2

The study population for Naval Shipyard #2 (NS2) consisted of 343,355 daily and 31,532 annual radiation records, which characterized the exposures of 4,525 workers (Table 2.1). At NS2, nearly all of the workers in the study population were white. Similar to NS1, the highest level of education for two-thirds of the workers was high school graduate. Approximately half of the workers were born prior to 1950.

The NS2 radiation dataset spans from 1975 to 2005, and there were 98 unique job titles on which radiation measurements were collected (Table 2.2*)*. Of the 343,355 daily measurements, 44.0% had a value of 0 mrem. The population mean exposure level was 20.9 mrem and the population median exposure level was 1.0 mrem. Of the 31,532 annual measurements, 15.9% had a value of 0 mrem. The annual population mean exposure level was 222.9 mrem and the median exposure level was 62.0 mrem. Compared to NS1, the population exposure levels at NS2 appeared to be higher overall, with lower percentages of zero values. On average, nearly 11 daily measurements were collected on a given worker during a one-year period.

2.1.5. Summary of Work Population at Shipyard #3

The study population at Naval Shipyard #3 (NS3) consisted of 290,394 daily and 25,081 annual radiation records, which characterized the exposures of 3,216 workers (Table 2.1). The race of workers at NS3 was more diverse, with more than 25% of the workers reported as Japanese and only 18.5% reported as white. The highest level of education for close to half of the workers was high school graduate; compared to NS1 and NS2, a higher percentage of workers completed a terminal occupational program or some college.

The NS3 radiation dataset spans from 1976 to 2005, and there were 87 unique job titles on which radiation measurements were collected (Table 2.2*)*. Of the 290,394 daily measurements collected, 62.8% had a value of 0 mrem. The population mean exposure level was 13.8 mrem and the population median exposure level was 0.0 mrem. Of the 25,081 annual measurements, 27.7% had a value of 0 mrem. The annual population mean exposure level was 158.2 mrem and the median exposure level was 21.0 mrem. The mean and median exposure levels for NS3 were closer in value to those of NS1 as compared to those of NS2. On average, approximately 11 daily measurements were collected on a given worker during a one year period.

Table 2.1. Summary of shipyard worker demographics

Shipyard	**NS1**		**NS2**		**NS3**	
Total size of work population	6,105		4,525		3,132	
Variable	No.	*%*	No.	*%*	No.	*%*
Race						
Black, not of Hispanic origin (%)	1617	*26.5*	25	*<1.0*	29	*<1.0*
Chinese	0	*0*	0	*0*	204	*6.5*
Filipino	0	*0*	0	*0*	488	*15.6*
Guamanian	0	*0*	0	*0*	14	*<1.0*
Hawaiian	0	*0*	0	*0*	347	*11.1*
Hispanic	51	*1.0*	10	*<1.0*	62	*2.0*
Japanese	0	*0*	0	*0*	847	*27.0*
White, not of Hispanic origin	4342	*71.0*	4471	*98.8*	578	*18.5*
Unknown	0	*0*	0	*0*	429	*13.7*
Other	95	*1.5*	19	*<1.0*	134	*4.3*
Highest Level of Education						
Some high school	272	*4.5*	267	*5.9*	117	*3.7*
High school graduate	3938	*64.5*	3005	*66.4*	1401	*44.7*
Completion of terminal occupational program	855	*14.0*	524	*11.6*	460	*14.7*
Some college	658	*10.8*	284	*6.3*	540	*17.2*
Associate Degree	118	*1.9*	137	*3.0*	296	*9.5*
Bachelor's Degree	109	*1.8*	68	*1.5*	133	*4.2*
Other	155	*2.5*	240	*5.3*	185	*6.0*
Birth Year						
Before 1950	1911	*31.3*	2315	*51.2*	1445	*46.1*
1950 or later	4194	*68.7*	2210	*48.8*	1687	*53.9*

Table 2.2. – Daily and annual radiation exposure levels by shipyard

Shipyard	NS1	NS2	NS3
Sample Collection Start Year	1975	1975	1976
Sample Collection End Year	2005	2005	2005
No. unique job titles	86	98	87
Daily Radiation Measurements			
No. total measurements	440463	343355	290394
No. measurements = 0 mrem	221743	151157	182395
% measurements = 0 mrem	50.3	44.0	62.8
Mean exposure level (mrem)	15.4	20.9	13.8
Median exposure level (mrem)	0.0	1.0	0.0
Peak exposure level (mrem)	8872	4784	3514
Annual Radiation Measurements			
No. total measurements	43462	31,532	25081
No. measurements = 0 mrem	10063	5,005	6952
% measurements = 0 mrem	23.2	15.9	27.7
Mean exposure level (mrem)	156.7	222.9	158.2
Median exposure level (mrem)	18.0	62.0	21.0
Peak exposure level (mrem)	8872	5086	3514
Avg. no. daily measurements per year	10.2	10.8	11.2

2.2. Radiation Exposure

2.2.1. Source of Exposure

The primary source of radiation exposure to the nuclear naval shipyard population is external gamma radiation emitted by activated corrosion products, primarily cobalt-60, deposited within the coolant systems of the reactor compartments of the submarines. Co^{60} emits two high-energy gamma rays and a low-energy beta particle for every radioactive decay (Daniels et al. 2004; Matanoski et al. 2008).

The nuclear workers in shipyards differ from more traditional radiation exposed professions in that their jobs are not directly related to working with a radiation source. Instead, they are exposed incidentally by handling radioactive materials while carrying out the normal tasks related to their trades (Matanoski et al. 2008). Thus, radiation exposure among shipyard

workers is only associated with specific trades, tasks, and locations within the ship. Nearly all the radiation dose in this worker population is attributed to tasks performed within the shielded reactor compartment onboard these nuclear-powered submarines (Daniels et al. 2004; Matanoski et al. 2008).

2.2.2. Exposure Monitoring

Between 1973 and 1974 (prior to the start of the exposure data used in this study), the shipyards converted to thermal luminescent dosimeters (TLDs) that were read daily (Matanoski et al. 2008). TLDs are based on the effect of thermoluminescence, a property exhibited by certain materials in which previously absorbed energy is re-emitted as light upon heating. In these dosimeters, absorption of energy from the radiation excites the atoms inside a crystal, which produces free electrons that remain trapped within the crystal's structure as excitation energy. Heating the crystal releases this energy as light, which can then be measured and is directly proportional to the radiation absorbed dose (Cember & Johnson, 2009).

The switch to TLDs resulted in a reduction in all annual doses due to the limitations of measuring low doses in the minimum detection range when cumulated doses are fractionated into daily readings, as opposed to the previously collected monthly readings using film badges (Matanoski et al. 2008). The radiation exposure concentrations used in these analyses represent the total effective dose equivalent (TEDE), which is the sum of external and internal exposures (Gollnick, 2006).

2.2.3. Daily and Annual Exposure Records

As mentioned above, two separate datasets of exposure records were constructed. In the first dataset, each entry represents a daily radiation level reading, in mrem. Each worker therefore has a separate entry for every dosimeter reading collected over the span of his employment (i.e., 10 entries for 10 daily readings). For the second dataset, all dosimeter readings for a given worker in a given year were summed, resulting in an annual, cumulative exposure for each worker and for each year of employment (i.e., 5 entries for 5 years).

Both the daily and annual exposure datasets were used for the analyses described in this dissertation. Use of one dataset over the other was based on the overall objective of the stated analysis. Daily readings represent a single exposure period, similar to the 8-hour time weighted average (8hr-TWA) commonly used by industrial hygienists when determining whether an exposure level exceeds an occupational exposure limit (OEL). The annual records represent a cumulative dose over a period of one year, similar to the cumulative dose levels used by epidemiologist and risk assessors when calculating risk estimates to characterize the association between an exposure and an outcome of interest.

2.3. Imputation Model

2.3.1. Introduction to Multiple Imputation

Imputation involves the substitution of an estimated value for one that is missing (Fielding et al. 2010). In multiple imputation, the set of possible values for the missing observations are based on the distribution of the data (Pigott, 2001). Estimates of missing values are obtained

by simulating random draws from a modeled distribution of the missing variable(s) given the observed variables. These estimated values, or imputes, are then substituted for the missing values, resulting in a complete dataset (Fielding et al. 2010; van Buuren, 2012). As suggested by the name multiple imputation, the process of selecting a random value from the distribution of each missing value is repeated many times, generating *m* multiple randomly different datasets (Enders, 2011; Fielding et al. 2010). The value *m* takes can vary and is generally at the discretion of the investigator, but research has suggested that even an *m* of less than 10 typically results in unbiased estimates (Rubin, 1987; Schafer & Olsen, 1998; von Hippel, 2005). The standard statistical analyses of interest are applied to each of the *m* imputed datasets individually, obtaining *m* estimates of the selected parameters. For each parameter, the *m* estimates are then combined using prescribed rules to produce one overall parameter estimate (Rubin, 1987). The estimated imputation variance is a combination of the conventional sampling variance (within-imputation variance) and the extra variance caused by the missing data (between-imputation variance) (Enders, 2011; van Buuren, 2012).

The purpose of multiple imputation is not to obtain individual values for each missing data point. Rather, by estimating multiple values for these missing data and thus creating complete datasets, important characteristics of the dataset as a whole are preserved and the parameter estimates, including means, variances, and linear regression coefficients, should be unbiased (Graham, 2009).

Multiple imputation has a number of important advantages. First, it restores the random variability in the missing data by creating imputed values, which are based on variables

correlated with the missing data (Fielding et al. 2010). This results in a dataset that maintains the overall variability in the population while preserving the relationships between variables. In addition, the process incorporates the uncertainty resulting from estimating missing data by creating multiple different version of the imputes (Fielding et al. 2010). Since multiple imputation has been gaining in popularity, there also now exist more software packages that support this technique, making it an easier option for data analysis.

One drawback to many of the model-based methods, like multiple imputation, is the required assumption of an MAR missing data mechanism, which can be difficult to prove (Pigott, 2001). Still, even when this assumption is violated, multiple imputation often emerges as a preferred approach (Graham, 2009). Multiple imputation is also more computationally intensive than some other methods (Sterne et al., 2009). With constant advancements in computational processing and software, however, this method is becoming more and more accessible (Fielding et al. 2010; van Buuren, 2012).

Given the advantages of multiple imputation, this method was considered to be a good candidate for addressing the missing exposure data often found in occupational cohorts. The importance of developing accurate exposure estimates despite the challenges that exist when missing data are present have already been highlighted. Using a multiple imputation approach has the potential to improve upon occupational exposure estimates.

2.3.2. Model Variables

The imputation model used when working with the daily radiation measurements included the following variables: sample collection year, sample collection quarter, job title, education level, birth year, and race. The model used when working with the annual radiation measurements was similar but did not include sample collection quarter, as the measurements were cumulative over a year.

2.3.3. Software

All imputations were performed using the *mice* package in R 3.1.0 (van Buuren & Groothuis-Oudshoorn, 2011). Since the exposure variable with missing data is continuous, imputations were generated by predictive mean matching. The number of imputations was set at $m = 5$.

2.3.4. Measures of Performance

To evaluate the performance of multiple imputation in characterizing the exposure profiles defined in each analysis, several measures of performance were used. These measures were selected following a review of publications with a similar objective (Ma et al. 2012; Mishra & Khare, 2014).

Bias: Bias is the difference between a pooled estimate of exposure as determined by multiple imputation, $\bar{\Theta}$, and the true parameter of exposure. For these analyses, the bias of the mean and median exposure estimates was examined.

Variance: The total variance of the multiply imputed datasets is estimated as a function of the average of the estimated sampling variance for each imputed estimate (the within-imputation variability) and a component that captures the imputation variability over the repetition of the imputation process (the between-imputation variability) (Ma et al. 2012; van Buuren, 2012).

Within-imputation variance, $\bar{U}$, is the variance resulting from taking a sample rather than observing the entire population. This is the conventional statistical measure of variability and is calculated using the formula:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\bar{U}_i$$

where $\bar{U}_i$ is the estimate of the variance of θ_m for each imputation and *m* is the number of imputations.

Between-imputation variance, B, is the variance caused by the fact that there are missing values in the sample. It is calculated using the formula:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\Theta}_m - \bar{\Theta})^2$$

The estimated *total variance* of the MI estimate is then calculated using the formula:

$$T = \bar{U} + (1 + \frac{1}{m})B$$

Standard error: Standard error is the square root of the total variance, $\sqrt{T}$. Standard error will be used to describe the total imputation variance and the true variance (van Buuren, 2012).

Interval Estimate: The 95% confidence interval was estimated for the pooled estimate of the mean exposure. Multiple imputation confidence intervals are constructed using the pooled estimate of interest, its standard error, and a critical value from the Student t distribution with v degrees of freedom (van Buuren, 2012):

$$100(1-\alpha)\% = \bar{\Theta} \pm t_{v,1-\alpha/2}\sqrt{T}$$

CHAPTER 3: ESTIMATING POPULATION-LEVEL EXPOSURES

3.1. Introduction

3.1.1. Specific Aims

The specific aims of this chapter are to understand common missing data patterns in occupational cohorts, examine the effect these patterns have on the ability to accurately characterize population-level exposures, and to test the performance of a multiple imputation approach in estimating such population exposures.

3.1.2. Focused Research Objectives

The analyses performed in this chapter are grouped into three sections and are summarized in Table 3.1. Each section examines the exposure data using a different pattern of missingness and attempts to answer more focused research questions related to the overall specific aims. In order to test the performance of the multiple imputation method, artificially missing exposure data were generated from each complete dataset. The method by which data were selected to be missing was varied for each analysis, with the goal being to generate missing data in ways that reflect real-world sampling scenarios. The overall objectives and challenges faced by industrial hygienists and epidemiologists when collecting and/or reviewing exposure data were considered.

3.1.2.1. Random Selection

In the first section, missing exposure data are selected randomly from the overall study population. This section explores the ability to estimate population exposures from a small,

randomly selected portion of exposure measurements. The analyses in this section address the following research objectives:

- Test the performance of a multiple imputation approach for addressing missing exposure data when the data are randomly selected to be missing
- Test the performance of a multiple imputation approach when the proportion of randomly selected missing exposure data is large (up to 99.99% missing)

3.1.2.2. Selection by Sampling Collection Date

In the second section, exposure data are selected to be missing based on the collection date of the radiation measurements. Exposure measurements collected in more recent years will be used to characterize earlier, retrospective exposures, the majority of which have been artificially assigned as missing. The analyses in this section address the following research objectives:

- Test the ability of multiple imputation to characterize exposure estimates for work populations with incomplete historical measurement data
- Compare the performance of multiple imputation when three varying subsets of exposure data are made available for use in the imputation process

3.1.2.3. Selection by Job Title

In the third section, exposure data are selected to be missing based on the job title of the worker population. Plausible industrial hygiene sampling plans will be simulated and the results will be used to characterize population-level exposures. The analyses in this section address the following research objectives:

- Test the ability of multiple imputation to estimate exposures for a large work population based on the selected availability of measurement data by job title
- Compare the performance of multiple imputation when various plausible industrial hygiene sampling plans are simulated

Table 3.1 Summary of analyses completed in Chapter 3

Section	Missing Data Pattern	Sub-analysis	Exposure Data Used
3.1 Random Selection	By random	50-99.99% missing	Daily
3.2 Selection by Sample Collection Date	By sample collection date (Early v. Recent Years)	*Full* – using all recent year data	Annual
		Non-zero – using all recent year data with values > 0 mrem	Annual
		Bin – using a preselected subset of recent year data categorized into 6 bins	Annual
		HML – using a preselected subset of recent year data categorized into 3 bins	Annual
3.3 Selection by Job Title	By job title of work population	*Equal Percentage* – equal percentage sampled from each job	Daily
		Prior exposure – oversampled by highest prior exposures	Annual
		Related exposure – oversampled based on asbestos exposure rankings	Annual
		Published literature – oversampled based on published leukemia OR	Annual

3.2. Methods

3.2.1. Random Selection

In occupational cohorts, the percentage of measurements that are missing or uncollected can be very high. It is not uncommon for an industrial hygienist to be faced with characterizing the exposures of an entire work population based on samples collected on 10%, or less, of the

worker population. It is not clear, however, how well multiple imputation will perform at such high percentages of missing occupational exposure data.

In this first section, missing exposure data were selected randomly from the overall study population. This section explores the ability to estimate population exposures from a small, randomly selected portion of exposure measurements. The analyses in this section address the following research objectives:

- Test the performance of a multiple imputation approach for addressing missing exposure data when the data are randomly selected to be missing
- Test the performance of a multiple imputation approach when the proportion of randomly selected missing exposure data is large (up to 99.99% missing)

The percentage of missing data artificially generated using the daily exposures dataset was varied at increments ranging from 50.0% to 99.99% missing (Table 3.2). The exposure data designated as "missing" were randomly selected; the selection was not based on the value of any other variable in the dataset or on the expected exposure levels. Seven separate scenarios were examined in which the percentage of missing data varied. The exposure data that were not classified as "missing" were considered to have been "sampled" and remained in the dataset. Each scenario was applied to the dataset of daily measurements for each of the three shipyards. Table 3.2 illustrates the number of daily exposure measurements that were designated as "missing" and "sampled." For instance, when 99.99% of the daily exposure data at NS1 were assigned to be missing, only 44 measurements were available for the

imputation process. Those exposure measurements that were designated as "sampled" were then used to impute the missing exposure data.

Table 3.2. Number and percentage of daily measurements assigned "missing" and "sampled" by percent missing and shipyard

Shipyard	NS1 (total n = 440,463)		NS2 (total n = 343,355)		NS3 (total n = 290,394)	
% Missing	*No. meas. assigned "missing"*	*No. meas. assigned "sampled"*	*No. meas. assigned "missing"*	*No. meas. assigned "sampled"*	*No. meas. assigned "missing"*	*No. meas. assigned "sampled"*
50.0	219,959	220,504	171,418	171,937	144,924	145,470
90.0	396,303	44,160	308,958	34,397	261,341	29,053
95.0	418,020	22,443	325,887	17,468	275,626	14,768
99.0	436,007	4,456	339,912	3,443	287,469	2,925
99.90	440,036	427	343,031	324	290,119	275
99.95	440,247	216	343,193	162	290,261	133
99.99	440,419	44	343,322	33	290,371	23

3.2.2. Selection by Sample Collection Date

When conducting an exposure assessment or epidemiology study, one of the variables that will likely most influence the availability of measurement data is the time period of exposure. Ideally, contemporaneous exposure measurement data will exist for the time period of interest to the study; however, given the limited availability of exposure data in many occupational studies, this often might not be the case. Instead, exposures may have to be estimated using sampling data that was partially, or completely, collected during a different time period. In many of these scenarios, the availability of exposure data decreases as the time period of interest grows earlier. The epidemiologist or exposure assessor may then be tasked with estimating population exposure levels using limited historical exposure data, which can make retrospective exposure assessments challenging (Symanski et al. 1998a). In such cases, historical exposure data may be supplemented with more recently collected

measurements, which do not always accurately reflect the exposure levels of previous time periods. An objective of this section therefore is to examine how the accuracy of population-level exposure estimates is affected by the use of exposure data from multiple time periods and the impacts this may have on the performance of a multiple imputation approach.

In the second section, exposure data are selected to be missing based on the collection date of the radiation measurements. Exposure measurements collected in more recent years will be used to characterize earlier, retrospective exposures, the majority of which have been artificially assigned as missing. The analyses in this section address the following research objectives:

- Test the ability of multiple imputation to characterize exposure estimates for work populations with incomplete historical measurement data
- Compare the performance of multiple imputation when three varying subsets of exposure data are made available for use in the imputation process

The annual exposure data for each yard were stratified into two time periods based on the sample collection date: "early" data (exposure readings with a sample collection date prior to 1990) and "recent" data (readings with a sample collection date of 1990 or later). Table 3.3 stratifies the annual exposure data by sample collection date. The recent year exposure data were observed to have a higher percentage of 0 mrem values and a lower mean and median value as compared to the early year data. This is not surprising, as occupational exposure levels are known to often decrease over time (Symanski et al. 1998b).

To test the performance of multiple imputation, a percentage of the early year data was artificially designated as "missing." The remaining early year data were then combined with the recent year data to impute the missing exposures. Four separate analyses were performed using the recent year data, which are described in detail in the next sections. In the first analysis, all of the exposure data within the recent year group were used in the imputations (Full Analysis). In the second analysis, only those measurements in the recent year group with exposure levels greater than 0 mrem were included (Non-zero Analysis). In the third analysis, a subset of the recent year group was created based on the stratification of the exposure measurements into six bins (Bin analysis). Finally, the fourth analysis also utilized a subset of the recent year group, this time based on the stratification of the exposure measurements into high, medium, and low bins (HML analysis). The same set of missing early year exposure data were used in each analysis to allow for comparisons.

Table 3.3. Stratification of annual datasets by sample collection date

Early *(sample collection date prior to 1990)*				
Shipyard	**Sample Size**	**Mean (mrem)**	**Median (mrem)**	**% 0 mrem**
NS1	16597	197.9	40.0	17.4
NS2	16423	283.6	104.0	11.6
NS3	11649	210.5	46.0	16.1
Recent *(sample collection date of 1990 or later)*				
Shipyard	**Sample Size**	**Mean (mrem)**	**Median (mrem)**	**% 0 mrem**
NS1	26865	131.3	11.0	26.7
NS2	15109	156.9	33.0	20.5
NS3	13432	112.8	7.0	37.8

3.2.2.1. Full Analysis

In the Full Analysis, all of the recent year data were used to impute the missing early year exposure data. The missing data in the early year group were generated such that 90% of the data were randomly assigned as missing. This analysis tests the hypothesis that the distribution of exposures in the recent year group is similar to that of the early year group and that using recent year data to impute early year data, without making any adjustments, will therefore result in accurate estimates of the population exposure levels.

3.2.2.2. Non-zero Analysis

In the Non-zero Analysis, a subset of the recent year exposure data was created such that all exposure values of 0 mrem were removed from the dataset. This subset was then used to impute the same 90% randomly missing early year exposures as in the previous analysis. Removal of the zero values is a response to both the uncertainty in the true exposure level of these measurements, which fall below the limit of detection, and the higher proportion of low-level exposures observed in the recent year group. This analysis tests the hypothesis that using a subset of the recent year data with the lowest exposure values removed will more accurately estimate the population exposure levels.

3.2.2.3. Bin Analysis

In the Bin Analysis, a subset of the recent year exposure data was created based on the stratification of exposure measurements into six bins. First, the original early year exposure data (without any missing exposures) and the recent year data were each stratified into six bins, based on exposure level (Table 3.4). The six bins were: 0 to <1 mrem; 1 to <5 mrem; 5

to <10 mrem; 10<100 mrem; 100 to <1000 mrem; and ≥ 1000 mrem. As Table 3.4 shows, a general pattern could be observed. A greater percentage of the recent year exposure data, as compared to the early year data, were assigned to the first bin for all three yards. Similarly, a greater percentage of the early year exposure data, as compared to the recent year data, were assigned to the fifth and sixth bins for all three yards. This stratification helps to illustrate why the early year mean and median exposure levels were higher than the recent year levels as shown in Table 3.3.

A subset of the recent year exposure data was then created that mirrored the proportions of early year exposure measurements placed in each bin (Table 3.4, third column). This subset was then used to impute the same 90% randomly missing early year exposures as in the two previous analyses. By creating this subset, the recent year data now more accurately represents the exposure profile of the early year data. This analysis therefore tests the hypothesis that using a subset of recent year that more closely reflects the frequency of exposure levels observed in the early year group will more accurately estimate the population exposure levels.

3.2.2.4. HML Analysis

In the HML analysis, a subset of the recent year exposure data was crated based on the stratification of exposure measurements into high, medium, and low bins. Similar to the bin analysis, the original early year exposure data and the recent year data were each stratified into three bins, based on exposure level (Table 3.5). The *low* exposure category was defined as 0 to <5 mrem; the *medium* exposure category was defined as 5 to <100 mrem; and the

high exposure category was defined as ≥100 mrem. As Table 3.5 shows, a greater percentage of the recent year exposure data, as compared to the early year data, were assigned to the first (*low*) bin; similarly, a greater percentage of the early year exposure data, as compared to the recent year data, were assigned to the third (*high*) bin.

A subset of the recent year exposure data was then created that mirrored the proportions of early year exposure measurements assigned to each bin *(*Table 3.5, third column). This subset was then used to impute the same 90% randomly missing early year exposures as in the three previous analyses. By creating the subset in this manner, the recent year data now more accurately represents the exposure profile of the early year data; furthermore, the subset was defined using categories that may be more accessible compared to the previous analysis. Use of a qualitative scale for defining exposure levels is a very common technique for assessing exposure, particularly when quantitative exposure data are scarce. Furthermore, the use of only three exposure bins, as opposed to the six bins used in the previous analysis, allows for an easier time assigning exposures and increases the probability that the correct exposure bin will be chosen. If quantitative exposure data are limited, or unavailable, worker exposure levels can be assigned to a bin based on the professional judgments of an expert in the field. Like the previous analysis, this analysis therefore tests the hypothesis that using a subset of recent year that more closely reflects the frequency of exposure levels observed in the early year group will more accurately estimate the population exposure levels. However, the stratifications created in this analysis may be considered a more feasible approach.

Table 3.4. Comparison of early v. recent year groups stratified into six exposure bins

	NS1 Early			**NS1 Recent: original**			**NS1 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
0 to <1	2892	17.4	17.4	7171	26.7	26.7	3132	17.4	17.4
1 to <5	2075	12.5	29.9	4222	15.7	42.4	2250	12.5	29.9
5 to <10	918	5.5	35.4	1734	6.5	48.9	990	5.5	35.4
10 to <100	4434	26.7	62.1	6720	25.0	73.9	4806	26.7	62.1
100 to <1000	5528	33.3	95.4	6177	23.0	96.9	5994	33.3	95.4
≥1000	750	4.6	100	841	3.1	100	828	4.6	100
Total	**16597**	**--**	**100**	**26865**	**--**	**100**	**18000**	**--**	**100**
	NS2 Early			**NS2 Recent: original**			**NS2 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
0 to <1	1904	11.6	11.6	3101	20.5	20.5	464	11.6	11.6
1 to <5	1573	9.6	21.2	1884	12.5	33.0	384	9.6	21.2
5 to <10	840	5.1	26.3	874	5.8	38.8	204	5.1	26.3
10 to <100	3798	23.1	49.4	3731	24.7	63.5	924	23.1	49.4
100 to <1000	7418	45.2	94.6	5290	35.0	98.5	1808	45.2	94.6
>1000	890	5.4	100	229	1.5	100	216	5.4	100
Total	**16423**	**--**	**100**	**15109**	**--**	**100**	**4000**	**--**	**100**
	NS3 Early			**NS3 Recent: original**			**NS3 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
0 to <1	1878	16.1	16.1	5074	37.8	37.8	548	16.1	16.1
1 to <5	1439	12.4	28.5	1357	10.1	47.9	423	12.4	28.6
5 to <10	641	5.5	34	551	4.1	52	188	5.5	34.1
10 to <100	3143	27.0	61	2823	21.0	73	918	27.0	61.1
100 to <1000	4024	34.5	95.5	3457	25.7	98.7	1173	34.5	95.6
>1000	524	4.4	100	170	1.3	100	150	4.4	100
Total	**11649**	**--**	**100**	**13432**	**--**	**100**	**3400**	**--**	**100**

Table 3.5. Comparison of early v. recent year groups stratified into High, Medium, and Low exposure bins

	NS1 Early			**NS1 Recent: original**			**NS1 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
Low (0 to <5)	4967	29.9	29.9	11393	42.4	42.4	4760	30.0	30.0
Medium (5 to <100)	5352	32.2	62.1	8454	31.5	73.9	5110	32.2	62.2
High (≥100)	6278	37.9	100	7018	26.1	100	6000	37.8	100
Total	**16597**	**--**	**100**	**26865**	**--**	**100**	**15870**	**--**	**100**
	NS2 Early			**NS2 Recent: original**			**NS2 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
Low (0 to <5)	3477	21.2	21.2	4985	33.0	33.0	2120	21.2	21.2
Medium (5 to <100)	4638	28.2	49.4	4605	30.5	63.5	2820	28.2	49.4
High (≥100)	8308	50.6	100	5519	36.5	100	5060	50.6	100
Total	**16423**	**--**	**100**	**26865**	**--**	**100**	**10000**	**--**	**100**
	NS3 Early			**NS3 Recent: original**			**NS3 Recent: Subset**		
Exposure Bins (mrem)	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*	*Frequency*	*Relative %*	*Cumulative %*
Low (0 to <5)	3317	28.5	28.5	6431	47.9	47.9	2562	28.5	28.5
Medium (5 to <100)	3784	32.5	61.0	3374	25.1	73.0	2922	32.5	61.0
High (≥100)	4548	39.0	100	3627	27.0	100	3516	39.0	100
Total	**11649**	**--**	**100**	**13432**	**--**	**100**	**9000**	**--**	**100**

3.2.3. Selection by Job Title

Another variable that often influences the availability of measurement data is the job title of the worker population. Industrial hygienists, when designing a sampling plan, most often consider job title or work task when determining how to divide the allotted number of samples between workers. As discussed previously, hygienists are almost always limited in the number of samples they can collect and will need to make decisions based on the potential for overexposure within each job. There are a number of ways a hygienist can design a sampling plan and the resulting estimates of the population exposure levels will be affected by how many samples were collected and on which workers. This influence on population exposure levels is an important consideration both for industrial hygienists, as they move forward with their recommendations, and for epidemiologists using the exposure data for their own analyses.

This section will examine the influence various plausible sampling plans have on the ability of multiple imputation to estimate population exposure levels. Exposure data are selected to be missing based on the job title of the worker population. Plausible industrial hygiene sampling plans will be simulated and the results will be used to characterize population-level exposures. The analyses in this section address the following research objectives:

- Test the ability of multiple imputation to characterize exposure estimates for a large work population based on the selected availability of measurement data by job title
- Compare the performance of multiple imputation when various plausible industrial hygiene sampling plans are simulated

For these analyses, missing data are generated in such a way as to reflect real-world industrial hygiene sampling plans. Broadly, two general sampling plans were considered: one in which job titles are evenly sampled and one in which certain job titles were oversampled due to some knowledge of expected exposure levels. Several possible sampling plan scenarios are investigated in this section and are described below.

3.2.3.1. Even Percentage Sampling

If an industrial hygienist has no *a priori* knowledge regarding expected exposures levels within each job title, they may choose to design a sampling plan in which an even number, or even percentage, of measurements are collected from each relevant job title. The goal of such a sampling plan may be two-fold: to estimate the population-level exposure profile, and to accurately characterize the exposure levels of each monitored job title. Collecting an even percentage of samples from each job may allow the hygienist the ability to better understand exposure within and between job titles. However, the number of samples that can feasibly be collected from each job will certainly be limited. In this section, the ability to accurately characterize the population exposure profile from an even, but limited, percentage of samples collected from each job title was investigated.

Five separate analyses were performed using the daily measurements datasets in which the equal percentage of radiation measurements selected from each job title was set at 50%, 20%, 10%, 5%, and 1%. The selected measurements were designated as "sampled;" the remaining measurements were designated as "not sampled" and were artificially set to be missing. The "sampled" measurements were then used to impute the missing, "not sampled"

measurements. A summary of the five analyses is provided in Table 3.6. These analyses test the hypothesis that designing a sampling plan in which an equal percentage of exposure measurements are collected from each job title will allow for an accurate characterization of the population-level exposures.

Table 3.6. Number of "sampled" exposure measurements by percent sampled per job title

Shipyard	50% Sampled	20% Sampled	10% Sampled	5% Sampled	1% Sampled
NS1	219856	88272	44099	22112	4399
NS2	171461	69006	34719	17378	3470
NS3	145088	58221	29167	14674	2976

3.2.3.2. Unequal Sampling

The remaining analyses assume that the industrial hygienist has some *a priori* knowledge regarding expected exposure levels within each job title and thus designs a sampling plan accordingly. The sampling plan may be influenced by various sources of exposure information, including prior exposure levels, professional judgments, surveys and questionnaires, information on related exposures, or the published literature. In this section, three different potential sources of exposure information will be utilized to simulate hypothetical sampling plans: prior exposures, related exposures, and published literature. Again, a measurement is considered "sampled" if it remains available in the dataset; measurements that are artificially assigned to be missing are considered "not sampled." The different sampling plans will be examined for how well each can accuracy characterize the population exposure levels through a multiple imputation approach.

3.2.3.2.1. Prior Exposures

When designing a sampling plan, the industrial hygienist may have access to historical exposure data for the same jobs and/or work population. Although caution should be employed, prior data can help to identify those jobs or workers with the potential for high exposure levels through analysis of historical exposure patterns. Those jobs identified as having high prior exposures then become a priority for the current sampling plan. If the hygienist is only able to collect a limited number (or percentage) of measurements from the workforce, those jobs with known high previous exposures might then be oversampled.

In this section, the population exposure was estimated using a subset of the annual exposure data of NS1 in which jobs were oversampled based on prior exposure levels. High exposure levels were determined both by highest mean exposure and highest peak exposure, as each metric could be of interest to the hygienist.

A total of eight analyses were performed, all using the annual data from NS1. For the first four analyses, a sampling plan was created for the year 1990 using exposure data from the previous decade (1979-1989). The 73 unique job titles employed over that decade were first ranked from highest to lowest mean exposure over the ten-year period; the top 10% (n=7) of jobs were identified (Table 3.7*).* However, some of the jobs with the highest mean exposures from 1979-1989 did not appear in the dataset in 1990; thus, the selection continued until the top seven jobs that were also listed in 1990 had been identified. The seven jobs were: machine tool operating, insulating, boilermaking, fabric working, miscellaneous industrial equipment maintenance, marine machinery mechanic, and welding. These final seven jobs

were then assigned to be oversampled in the 1990 sampling plan, meaning all measurements were used and none were designated as missing. The measurements for all the remaining jobs were pooled together and 50% were designated as missing.

This analysis was then repeated, only this time the top 20% of jobs (n=14) were identified and oversampled. In addition to the seven jobs listed above, an additional seven were included. These jobs were: shipfitting, upholstering, miscellaneous general maintenance and operations work, painting, metal forging, sheet metal mechanic, and pipefitting. The measurements for all the remaining jobs were pooled together and 50% were designated as missing.

These two investigations were then repeated using peak exposure levels instead of mean exposures (Table 3.8). Many of the job titles selected for highest peak exposures were also selected for highest mean exposures over that same time period. The seven job titles used in the first analysis (when 10% of jobs were selected) were: insulating, rigging, miscellaneous industrial equipment maintenance, sheet metal mechanic, machining, boilermaking, and pipefitting. The additional seven jobs used in the second analysis (when 20% of jobs were selected) were: marine machinery mechanic, fabric working, welding, miscellaneous general maintenance and operations work, heavy mobile equipment mechanic, electrician, and electronics mechanic. Again, the measurements for all the remaining jobs were pooled together and 50% were designated as missing.

For the remaining four analyses, a sampling plan was created for the year 2000 using exposure data from the previous decade (1989-1999). The same four investigations as described above were repeated but using mean and peak exposure data from 1989-1999. The jobs with the highest mean and peak exposures over 1989-1999 are shown in Table 3.9 and Table 3.10. Compared to the 1979-1989 time period, both the mean and peak exposures from 1989 through 1999 were lower, suggesting a decline in exposure levels over time. However, many of the same job titles were still selected for highest mean and peak exposures, suggesting that while the exposure levels changed with time, the job titles with the highest relative exposures remained consistent.

The eight separate analyses are summarized in Table 3.11. These analysis test the hypothesis that, in the absence of contemporary exposure measurements, using prior exposure data from the same work location and job population can assist in developing a sampling plan that will accurately characterize the exposure profile of the work population.

Table 3.7. Mean exposure rankings in 1990 by job title based on annual data from 1979-1989

Mean Exposure Rank	Job Title	Mean Exposure (mrem)	Number of Measurements Available in 1990
1	Canvas Working	306.0	NA*
2	Boat Repairing	232.0	NA
3	Machine Tool Operating	224.3	2
4	Insulating	212.0	200
5	Pipe Covering	195.7	NA
6	Industrial Equipment Mechanic	184.0	NA
7	Boilermaking	178.3	544
8	Fabric Working	178.0	147
9	Miscellaneous Marine Maintenance	172.9	NA
10	Miscellaneous Industrial Equipment Maintenance	156.6	54
11	Equipment Cleaning	127.9	NA
12	Laboring	126.4	NA
13	Electronics Mechanic	125.6	NA
14	Marine Machinery Mechanic	116.4	781
15	Electronic Integrated Systems Mechanic	103.3	NA
16	Welding	103.1	351
17	Shipfitting	100.6	365
18	Electrical Equipment Repairer	97.3	NA
19	Rigging	81.5	NA
20	Upholstering	80.3	2
21	Miscellaneous General Maintenance and Operations Work	78.5	761
22	Coppersmithing	75.3	NA
23	Painting	72.0	141
24	Metal Forging	67.2	5
25	Electrical Inspecting Marine	62.8	NA
26	Sheet Metal Mechanic	62.0	247
27	Marine Maintenance Insp	60.5	NA
28	Pipefitting	59.1	1200

*NA indicates that the job title did not appear in the dataset in 1990

Table 3.8. Peak exposure rankings in 1990 by job title based on annual data from 1979-1989

Peak Exposure Rank	Job Title	Peak Exposure (mrem)	Number of Measurements Available in 1990
1	Insulating	8872	200
2	Rigging	3676	718
3	Miscellaneous Industrial Equipment Maintenance	1858	54
4	Sheet Metal mechanic	1818	247
5	Machining	1800	287
6	Boilermaking	1793	544
7	Pipefitting	1782	1200
8	Marine Machinery Mechanic	1740	781
9	Fabric Working	1631	147
10	Industrial Equipment Mechanic	1596	NA*
11	Welding	1502	351
12	Miscellaneous General Maintenance and Operations Work	1405	761
13	Heavy Mobile Equipment Mechanic	1159	45
14	Electrician	1114	544
15	Electronics Mechanic	1078	126

* NA indicates that the job title did not appear in the dataset in 1990

Table 3.9. Mean exposure rankings in 2000 by job title based on annual data from 1989-1999

Mean Exposure Rank	Job Title	Mean Exposure (mrem)	Number of Measurements Available in 2000
1	Machine Tool Operating	161.4	NA*
2	Insulating	77.1	1546
3	Fabric Working	56	911
4	Boilermaking	46.4	3220
5	Coppersmithing	45.3	NA
6	Boiler Plant Operating	34.7	NA
7	Metal Forging	34.2	111
8	Shipfitting	31.8	3961
9	Marine Machinery Mechanic	31.3	3396
10	Welding	26.6	2745
11	Miscellaneous General Maintenance and Operations Work	26	2253
12	Electronic Industrial Controls Mechanic	24.4	22
13	Machining	24.2	1016
14	Painting	21.2	1877
15	Pipefitting	18.7	6400
16	Sheet Metal Mechanic	18	1343
17	Air Conditioning Equipment Mechanic	16.8	595

* NA indicates that the job title did not appear in the dataset in 2000

Table 3.10. Peak exposure rankings in 2000 by job title based on annual data from 1989-1999

Peak Exposure Rank	Job Title	Peak Exposure (mrem)	Number of Measurements Available in 2000
1	Marine Machinery Mechanic	5479	3396
2	Boilermaking	1709	3220
3	Machining	1687	1016
4	Fabric Working	1588	911
5	Pipefitting	1494	6400
6	Shipfitting	1487	3961
7	Sheet Metal mechanic	1447	1343
8	Welding	1432	2745
9	Rigging	1374	4483
10	Miscellaneous General Maintenance and Operations Work	1352	2253
11	Insulating	1340	1546
12	Electrician	1266	3528
13	Machine Tool Operating	1000	NA*
14	Painting	863	1877
15	Air Conditioning Equipment Mechanic	837	595

*NA indicates that the job title did not appear in the dataset in 2000

Table 3.11. Summary of analyses using prior exposure data

Analysis	Year of Sampling Plan	Exposure Metric	No. of Jobs Oversampled	Number of Meas. in Oversampled Job Group (% of total meas.)
1	1990	Mean	7	2079 (28.1)
2	1990	Mean	14	4830 (65.2)
3	1990	Peak	7	3250 (43.9)
4	1990	Peak	14	6005 (81.1)
5	2000	Mean	7	15890 (37.3)
6	2000	Mean	14	29396 (69.1)
7	2000	Peak	7	20247 (47.6)
8	2000	Peak	14	37274 (87.6)

3.2.3.2.2. Related Exposures

If prior exposure data are not available, the industrial hygienist may turn to alternative sources of exposure information to assist in developing a sampling plan. One potential source of information may come from other exposures to the same work population that may be related to the exposure of interest. For example, in the shipyard worker population, asbestos is another common exposure of concern. Historically, asbestos was used ubiquitously on submarines as insulation on pipes and boilers; asbestos exposures were observed in many jobs in the fields of pipefitting, plumbing, woodworking, shipfitting, and welding (Zaebst et al. 2009). Much of the work that resulted in asbestos exposures is also of concern for exposure to radiation. Given the work they perform, many of the skilled trades workers who are at risk for high asbestos exposures may then also be at risk for high exposures to radiation. Information on the asbestos exposure levels of the work population to be monitored may therefore assist the industrial hygienist in making decisions regarding a sampling plan for radiation exposure.

As part of the epidemiology study on cancer risk of low-level radiation exposure that was described in Chapter 2, surveys of five shipyard industrial hygienists were conducted to assess asbestos exposure profiles for each job title listed in the study population (Matanoski et al., 2008; Correa-Villasenor, 1987). These surveys were used to construct a proxy indicator of relative intensity of asbestos exposure associated with each job. Four categories of asbestos exposure were defined:

Direct (4) – certain or probable exposure through direct handling of asbestos material

Indirect + Direct (3) – certain or probable indirect exposure and occasional direct exposure. Indirect exposure refers to the intermittent and usually occasional exposure to asbestos that workers whose jobs did not involve the use of asbestos may, nonetheless, have incurred as a result of being in the same general working environment where asbestos materials were being used

Indirect (2) – certain or probable indirect exposure

None (1) – minimal indirect and direct exposure to asbestos

Each job title in the study population was assigned an asbestos exposure category ranking of 1-4 through the survey of industrial hygienists. The job titles used in that research study are similar to those used in this dissertation and thus the asbestos category assigned to each job title was transferred to the corresponding job titles in this dissertation's study population.

The assumption being made in this section is that high asbestos exposures correspond to high radiation exposures. Applying the asbestos rankings to the job titles present in the annual datasets resulted in three job titles assigned to an asbestos exposure category of 4: insulating, heavy mobile equipment mechanic, and boilermaking. Four job titles were assigned to an asbestos exposure category of 3: welding, pipefitting, marine machinery mechanic, and electronics mechanics. In most cases, these job titles were shown to have mean and median radiation exposure levels that were near or above the population mean and median levels

(Table 3.12). The exceptions occurred in the job titles of heavy mobile equipment mechanic and electronics mechanic. In particular, the radiation exposure levels of heavy mobile equipment mechanics were quite low. Given the nature of the mechanics' work, which involves removing and replacing brake pads and linings, it would seem reasonable that they would have high asbestos exposures but low radiation exposures. This is a potential disadvantage of using other exposures as a proxy for the exposure of interest. To test the effect of such a scenario, each analysis will be performed twice – once with heavy mobile equipment mechanics included and once with them removed.

For the analyses, the jobs assigned an asbestos exposure category ranking of 3 or 4 were grouped together into a "High Rank" group. All remaining job titles were grouped together into a "Low Rank" group. A comparison of the mean and median radiation exposure levels for each group is shown in Table 3.13. The mean and median exposures were higher in the High Rank group in both cases, whether mechanics were included or not.

Similar to the analyses using prior exposure data, the jobs in the High Rank group were assigned to be oversampled, meaning all measurements were used and none were designated as missing. In the Low Rank group, 50% of the measurements were designated as missing. The remaining available measurements in the Low Rank group were then combined with the measurements from the High Rank group to impute the missing exposures. This analysis tests the hypothesis that, in the absence of measurement data on the exposure of interest, using a related exposure that can be reasonably assumed to have a similar exposure profile can assist in developing an accurate characterization of the population-level exposure profile.

Table 3.12. Annual radiation exposure levels of job titles assigned an asbestos exposure ranking of 3 or 4

Job Title	No. in Dataset (%)	Asbestos Ranking	Job-specific Mean (mrem)	Job-specific Median (mrem)
*NS1 (*Pop. *mean = 156.7 mrem, median= 18.0 mrem)*				
Boilermaking	2404 (5.5)	4	401.8	95.5
Heavy Mobile Equipment Mechanic	512 (1.2)	4	3.5	0.00
Insulating	1141 (2.6)	4	486.5	336.0
Electronics Mechanic	805 (1.9)	3	68.7	4.0
Marine Machinery Mechanic	3991 (9.2)	3	212.2	19.0
Pipefitting	7108 (16.4)	3	130.0	24.0
Welding	2387 (5.5)	3	167.3	40.0
NS2 (Pop. mean = 222.9 mrem, median = 62.0 mrem)				
Boilermaking	1 (0.0)	4	0.0	0.0
Heavy Mobile Equipment Mechanic	301 (1.0)	4	3.9	2.0
Insulating	883 (2.8)	4	480.0	181.0
Electronics Mechanic	624 (2.0)	3	66.9	5.0
Marine Machinery Mechanic	3167 (10.0)	3	306.8	130.0
Pipefitting	4466 (14.2)	3	203.1	78.0
Welding	1709 (5.4)	3	276.1	194.0
NS3 (Pop. mean = 158.2 mrem, median = 21.0 mrem)				
Boilermaking	254 (1.0)	4	123.6	30.5
Heavy Mobile Equipment Mechanic	259 (1.0)	4	12.7	0.0
Insulating	513 (2.0)	4	339.2	328.0
Electronics Mechanic	310 (1.2)	3	31.9	5.5
Marine Machinery Mechanic	3167 (12.6)	3	229.5	41.5
Pipefitting	2782 (11.1)	3	187.8	46.5
Welding	991 (4.0)	3	186.6	114.0

Table 3.13. Comparison of radiation exposure levels of high rank v. low rank group, with and without heavy mobile equipment mechanics

Shipyard	**NS1** *Pop. mean = 156.7 mrem* *Pop. median = 18.0 mrem*			**NS2** *Pop. mean = 222.9 mrem* *Pop. median = 62.0 mrem*			**NS3** *Pop. mean = 158.2 mrem* *Pop. median = 21.0 mrem*		
Group Based on Asbestos Exposure Category	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**
High Rank *with mechanics included*	18348 (42.2)	204.3	32.0	11151 (35.4)	252.7	105.0	9148 (36.5)	182.1	40.0
Low Rank	25114 (57.8)	121.9	13.0	20381 (64.6)	206.6	45.0	15933 (63.5)	144.4	14.0
High Rank *without mechanics included*	17836 (41.0)	210.1	37.0	10850 (34.4)	259.6	116.0	8738 (34.9)	190.3	50.0
Low Rank	25626 (59.0)	119.6	11.0	20682 (65.6)	203.7	42.0	16343 (65.2)	141.0	12.0

3.2.3.2.3. Published Literature

Finally, if no quantitative or qualitative exposure information for the specific work location of interest is available, the industrial hygienist may turn to the published literature for exposure information collected at similar facilities or on similar job titles. While there may be some concern about the generalizability of exposure data from one specific work site to other locations or time periods, etc., this source of exposure information may be an appropriate alternative when no others exist.

If raw exposure measurement data are not available, it may be possible to glean exposure information from risk estimates calculated by using the exposure of interest and a relevant biological outcome. A publication by Stern et al. that looked at the risk of leukemia by job ever held at a naval nuclear shipyard was used as the source of exposure information in this section (Sterne et al. 1986). The study population for this case-control analysis included male shipyard workers employed at one of the three shipyards used in this dissertation. The study period was from 1952 until 1977, which very slightly overlaps with the time period of this analysis. Cases were defined as all deceased persons by the end of 1980 for whom leukemia had been coded as an underlying or contributory cause of death. Risk factors of interest in the study were exposure to ionizing radiation and solvents. The publication included a table of the univariate analysis of leukemia by jobs and shops in which three or more cases ever worked (Table 3 in the original publication, reproduced as Table 3.14). Although the odds ratios were mostly not statistically significant, the corresponding jobs can be considered potentially highly exposed.

The job titles in this dissertation's study population that corresponded to those listed in Table 3.14 were identified (Table 3.15). These jobs were then grouped together in a "High Risk" group with a few adjustments. The job titles of carpenter, supervisor, and engineer were excluded from the group. Carpenters were only identified as being part of the study population at NS2 and thus were removed for comparison purposes. Supervisor was too broad of a title to identify an appropriate corresponding job, and engineer was not one of the job titles included in the selection criteria. Outside machinist and inside machinist were both represented by the job title machinist.

All remaining job titles were grouped together and categorized as the "Low Risk" group. As Table 3.16 shows, the Low Risk group had lower annual radiation mean and median exposure levels in two of the three shipyards. In NS1, the Low Risk group actually had higher annual mean and median radiation exposure levels, although the difference between the High and Low Risk group was small.

For the analysis, the High Risk group was assigned to be oversampled, meaning all annual measurements were used and none were designated as missing. In the Low Risk group, 50% of the measurements were designated as missing. The exposure measurements from the High Risk group, combined with the remaining 50% from the Low Risk group, were then used to impute the missing exposures. This analysis tests the hypothesis that, in the absence of available exposure data for the population of interest, using exposure information from a similar work population the published literature can assist in developing an accurate characterization of the population-level exposure profile.

Table 3.14. Univariate analysis of leukemia by jobs in which three or more cases ever worked ***(Table 3 from Stern et al. 1986)***

Job Ever Held	No. of exposed cases	OR for all leukemia	95% CI
Electrician	11	3.00	1.29-6.98
Carpenter	7	2.50	0.91-6.90
Supervisor	13	2.36	0.95-5.86
Welder	7	2.36	0.92-5.53
Sheet metal worker	4	2.14	0.64-7.19
Shipfitter	11	1.54	0.67-3.54
Engineer	6	1.40	0.53-3.70
Outside Machinist	6	0.91	0.38-2.22
Rigger	5	0.81	0.28-2.34
Pipefitter	5	0.70	0.27-1.84
Inside Machinist	10	0.54	0.24-1.22

Table 3.15. Summary of radiation exposure levels of job titles included in Stern et al. 1986 publication

Job Title	**No. in Dataset (%)**	**Job-specific Mean (mrem)**	**Job-specific Median (mrem)**
NS1 *(Pop. mean = 156.7 mrem, median= 18.0 mrem)*			
Carpentry	0 (0.0)	--	--
Electrician	4291 (9.9)	90.5	7.0
Machinists	1519 (3.5)	115.4	2.0
Pipefitting	7108 (16.4)	130.0	24.0
Rigging	3783 (8.7)	85.5	15.0
Sheet metal mechanic	1621 (3.7)	117.7	28.0
Shipfitting	2566 (5.9)	265.1	81.0
Welding	2387 (5.5)	167.3	40.0
NS2 *(Pop. mean = 222.9 mrem, median = 62.0 mrem)*			
Carpentry	34 (0.1)	1.3	0.5
Electrician	3124 (9.9)	188.6	40.0
Machinists	776 (2.5)	325.9	6.0
Pipefitting	4466 (14.2)	203.1	78.0
Rigging	1859 (5.9)	70.3	34.0
Sheet metal mechanic	619 (2.0)	160.6	72.0
Shipfitting	2327 (7.4)	379.7	345.0
Welding	1709 (5.4)	276.1	194.0
NS3 *(Pop. mean = 158.2 mrem, median = 21.0 mrem)*			
Carpentry	0 (0.0)	--	--
Electrician	2273 (9.1)	110.9	12.0
Machinists	1031 (4.1)	189.1	4.0
Pipefitting	3448 (13.7)	172.0	4.0
Rigging	1424 (5.7)	54.5	10.0
Sheet metal mechanic	901 (3.6)	151.6	51.0
Shipfitting	1813 (7.2)	375.7	234.0
Welding	1287 (5.1)	172.7	102.0

Table 3.16. Comparison of radiation exposure levels of high risk v. low risk group

Shipyard	**NS1** *Pop. mean = 156.7 mrem* *Pop. median = 18.0 mrem*			**NS2** *Pop. mean = 222.9 mrem* *Pop. median = 62.0 mrem*			**NS3** *Pop. mean = 158.2 mrem* *Pop. median = 21.0 mrem*		
Group Based on Leukemia Odds Ratio	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**	**No. in Dataset (%)**	**Mean (mrem)**	**Median (mrem)**
High Risk Group	23275 (53.6)	132.4	18.0	14880 (47.2)	224.1	78.0	12177 (48.6)	177.2	37.0
Low Risk Group	20187 (46.4)	184.7	19.0	16652 (52.8)	221.8	46.0	12904 (51.4)	140.2	10.0

3.2.4. Alternative Methods for Addressing Missing Data

While multiple imputation has gained popularity as a technique for addressing missing data, several alternative methods are still widely used (Pigott, 2001). Two of those methods, complete case analysis and mean substitution, were performed on a subset of the analyses completed in this chapter to allow for a comparison of the performances of all three approaches. As described in Chapter 1, complete case analysis includes only those subjects with no missing data, known as complete cases, resulting in a subset of the original study population that may potentially be biased (Demissie et al. 2003; Enders, 2011). In mean substitution, the missing data are filled-in with the mean value of the variable based on the available data, keeping the entire dataset intact but underestimating the variance (Pigott, 2001).

3.3. Results

3.3.1. Random Selection

Seven separate analyses were performed on each of the three shipyards' daily exposure datasets. The percentage of exposure data missing was increased in each subsequent analysis, from 50% missing to 99.99% missing.

Estimated Population Mean

Within each shipyard, the multiple imputation method performed reasonably well in estimating the population mean up through 99.90% of the data missing (Table 3.17). In all but one of the 15 trials that spanned 50-99.90% missing, the estimated population mean (MI mean) was less than 10 mrem from the true population mean; in all 15 trials, the MI mean

slightly overestimated the true mean. The confidence intervals for these trials were quite large and while they did appear to decrease with an increase in percentage of missing data for NS1, that pattern was not observed for the other two yards. When the highest percentage of exposure data were missing (99.95% and 99.99% missing), the multiple imputation approach did not perform as reliably, producing estimates of the mean that were inconsistent in size of the bias from the true population mean. The width of the confidence intervals also fluctuated significantly, from a largest width of 475 mrem to a smallest width of 61 mrem.

Estimated Population Median

The multiple imputation approach performed quite well in estimating the population median exposure (Table 3.17). The estimated population median (MI median) never differed more than 1.5 mrem from the true population median.

Estimated Imputation Variance

The within-imputation variances were much larger than the between-imputation variances (Table 3.17*)*. Within NS1, the imputation variance appeared to decrease as the percentage of missing data increased; however, the same pattern was not observed within NS2 or NS3. Until the percentage of missing data was set to the highest levels, the multiple imputation procedure maintained the true variance observed in the original datasets fairly well.

Table 3.17. Performance of MI by percent of missing data

% Missing	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (n = 440463, mean = 15.4 mrem, median = 0.0 mrem, SE = 78.4)						
50.0	1.2	(0, 178)	0.0	6.81E+03	6.2	82.6
90.0	1.4	(0, 179)	0.6	6.83E+03	8.2	82.7
95.0	1.2	(0, 173)	1.0	6.40E+03	6.9	80.0
99.0	0.0	(0, 157)	1.0	5.19E+03	6.3	72.1
99.90	1.4	(0, 144)	1.0	4.22E+03	16.4	65.1
99.95	17.9	(0, 190)	1.0	5.78E+03	499.5	79.9
99.99	-2.3	(0, 103)	0.8	2.10E+03	4.9	45.9
NS2 (n = 343355, mean = 20.9 mrem, median = 1.0 mrem, SE = 95.8)						
50.0	2.4	(0, 225)	0.0	1.05E+04	99.5	103.0
90.0	0.7	(0, 199)	0.0	8.15E+03	65.2	90.7
95.0	0.9	(0, 196)	0.0	7.87E+03	68.4	89.2
99.0	1.6	(0, 200)	0.0	8.06E+03	83.6	90.3
99.90	9.6	(0, 244)	0.4	1.14E+04	428.1	109.1
99.95	17.7	(0, 299)	0.2	1.66E+04	883.6	132.8
99.99	49.2	(0, 476)	2.0	3.64E+04	5444.2	207.1
NS3 (n = 290394, mean = 13.8 mrem, median = 0.0 mrem, SE = 89.4)						
50.0	0.8	(0, 190)	0.0	7.98E+03	0.1	89.3
90.0	0.5	(0, 185)	0.0	7.62E+03	0.8	87.3
95.0	1.7	(0, 195)	0.0	8.40E+03	2.2	91.6
99.0	4.5	(0, 205)	0.0	9.01E+03	35.5	95.1
99.90	10.7	(0, 236)	0.0	1.14E+04	164.4	107.8
99.95	21.1	(0, 303)	3.0	1.81E+04	544.7	136.9
99.99	-2.0	(0, 61)	0.0	6.18E+02	19.9	25.3

3.3.2. Selection by Sample Collection Date

Four separate analyses were performed on each of the three shipyards' annual exposure datasets. In each analysis, exposure data with a "recent" sample collection date (1990 or later) were used to impute missing exposure data with an "early" sample collection data (prior to 1990).

Full Analysis

In the "Full" analysis, all of the recent year data were used to impute the missing data. This resulted in underestimates of the true population mean for NS1 and NS2 (Table 3.18*)*. This is reasonable, as the missing early year exposure data were imputed using recent exposure data that were shown to have higher average exposure levels. The same pattern was not observed at NS3; however, the MI mean for NS3 was very close to the true mean. The estimates of the population median were underestimates of the true population median at all three yards, which again, seems reasonable.

Non-zero Analysis

In the "Non-zero" analysis, all of the recent year exposure data with a value of 0 mrem were removed from the dataset prior to imputation (Table 3.18). This resulted in MI mean estimates that were still underestimates of the true mean for NS1 and NS2 but closer in accuracy than those estimates observed in the Full Analysis. Again, this pattern was not observed within NS3, which is surprising given that the percentage of zeroes observed in the original datasets were comparable between the three yards. The MI median estimates were close to the true median for NS1 and NS2 but again, not for NS3.

Bin Analysis

In the "Bin" analysis, a subset of the recent year exposure data were selected that mirrored the proportions of early year exposure measurements placed in each of six bins (Table 3.18). This resulted in MI mean estimates that were reasonably close to the true population mean

for all three yards (within 10 mrem from the true value). The MI median estimates were also close to the true median for all three yards (within 5 mrem from the true value).

HML Analysis

In the "HML" analysis, a subset of the recent year exposure data were selected that mirrored the proportions of the early year exposure measurements placed in each of three bins (Table 3.18). This approach performed similarly as well in estimating the mean exposures as did the "Bin" analysis, with the exception of within NS2. The MI median estimates were again close to the true median for all three yards (within 5 mrem from the true value).

Estimated Imputation Variance

The within-imputation variances were much larger than the between-imputation variances in all trials (Table 3.18). The total variance produced by the multiple imputation procedure reflected the true variance observed in the original datasets for all four analyses.

Table 3.18. Performance of MI by recent year analysis (selection by sample collection year)

Recent Year Analysis	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (n = 43462, mean = 156.7 mrem, median = 18.0, SE = 314.2)						
Full	-18.5	(0, 767)	-5.2	8.56E+04	1.9	292.6
Non-zero	-2.1	(0, 758)	2.7	9.48E+04	0.9	307.8
Bin	-0.7	(0, 761)	0.6	9.56E+04	3.6	309.3
HML	-0.1	(0, 764)	0.4	9.61E+04	5.6	310.0
NS2 (n = 31532, mean = 222.9 mrem, median = 62.0, SE = 352.1)						
Full	-37.8	(0, 759)	-19.0	8.58E+04	4.3	292.9
Non-zero	-21.0	(0, 784)	-0.2	8.82E+04	10.1	297.0
Bin	-1.1	(0, 869)	4.4	1.09E+05	29.3	330.0
HML	-18.0	(0, 791)	3.3	8.93E+04	1.4	298.8
NS3 (n = 25081, mean = 158.2, median = 21.0, SE = 315.8)						
Full	1.3	(0, 796)	-5.0	1.05E+05	5.3	324.6
Non-zero	15.1	(0, 813)	10.2	1.06E+05	17.2	326.2
Bin	9.7	(0, 797)	3.6	1.03E+05	24.2	320.9
HML	7.6	(0, 787)	4.6	1.00E+05	10.8	317.0

3.3.3. Selection by Job Title

3.3.3.1. Equal Percentage Sampling

Five separate analyses were performed on each of the three shipyards' daily exposures dataset in which the percentage of radiation measurements drawn equally from each job was set at 50%, 20%, 10%, 5%, and 1%.

Estimated Population Mean

For all three shipyards, the estimates of the population mean were quite close to the true mean values (Table 3.19). For NS1, all of the MI mean estimates were within 2 mrem of the true value. The confidence intervals, while wide, stayed fairly constant in width as the percentage of measurements collected from each job decreased from 50% to 1%. For NS2, the MI mean estimates were again very close to the true value of the mean, with an increase

in bias observed when the percentage collected was set at 1%; the associated confidence interval was also wider than those observed at other percentage levels. For NS3, the MI mean estimates were also very close to the true value of the mean; similar to NS2, an increase in bias and confidence interval width were observed when the percentage of measurements collected was set at 1%.

Estimated Population Median

The estimates of the population median were exact estimates of, or close to, the true median for all five analyses at all three shipyards (Table 3.19). The observed bias was always less than 1.0 mrem.

Estimated Imputation Variance

The within-imputation variance was much larger than the between-imputation variance (Table 3.19). The estimates of the total imputation variance were overestimations of the true variance in all cases, although the estimates were still reasonably close. The variance did not differ greatly between analyses within a yard, with the possible exception of the 1% collected analysis for NS2.

Table 3.19. Performance of MI by percent of measurements collected per job title (equal percentage sampling)

% Collected Per Job	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (n = 440463, mean = 15.4 mrem, median = 0.0 mrem, SE = 78.4)						
50	1.3	(0, 177)	0.0	6.74E+03	7.0	82.1
20	0.7	(0, 182)	0.0	7.18E+03	4.3	84.7
10	1.5	(0, 182)	0.0	7.11E+03	8.6	84.4
5	1.4	(0, 183)	0.4	7.21E+03	8.9	85.0
1	1.2	(0, 172)	0.6	6.28E+03	12.1	79.4
NS2 (n = 343355, mean = 20.9, median = 1.0 mrem, SE = 95.8)						
50	2.3	(0, 221)	0.0	1.01E+04	98.2	101.1
20	1.2	(0, 204)	0.0	8.58E+03	74.6	93.1
10	1.0	(0, 208)	0.0	8.96E+03	71.0	95.1
5	1.1	(0, 217)	0.0	9.83E+03	73.3	99.6
1	5.9	(0, 290)	0.0	1.78E+04	196.0	134.3
NS3 (n = 290394, mean = 13.8 mrem, median = 0.0 mrem, SE= 89.4)						
50	0.9	(0, 197)	0.0	8.71E+03	0.2	93.3
20	0.9	(0, 190)	0.0	8.02E+03	0.3	89.5
10	1.9	(0, 200)	0.0	8.90E+03	7.9	94.4
5	1.5	(0, 184)	0.0	7.43E+03	1.9	86.2
1	2.8	(0, 212)	0.0	9.96E+03	7.2	99.8

3.3.3.2. Prior Exposures

Eight separate analyses were performed using the annual exposure dataset of NS1. In the first four analyses, a sampling plan was designed for the year 1990 using either the mean or peak exposure values from the previous ten years (1979-1989). In the remaining analyses, a sampling plan was designed for the year 2000 using either the mean or peak exposure values from the previous ten years (1989-1999).

Estimated Population Mean

The estimates of the population mean were very close to the true value of the population mean in both the 1990 and 2000 sampling plans (Table 3.20). Using either the mean or peak exposure values produced comparable estimates and confidence intervals.

Estimated Population Median

Similar to the population mean, the estimates of the population median were also fairly accurate (Table 3.20).

Estimated Imputation Variance

The within-imputation variance was greater than the between-imputation variance when designing a sampling plan for 1990; however, the two variances were quite similar when designing a sampling plan for 2000 (Table 3.20). This is likely a reflection of the higher percentages of zero and low values observed by the year 2000. In all eight analyses, the estimated total imputation variance was an overestimate of the true variance.

Table 3.20. Performance of MI by sampling plan year and exposure metric (prior exposures)

Trial	Exposure Metric (% included)	Mean Bias (mrem)	95% CI of Mean (mrem)	Median Bias (mrem	Within Variance	Between Variance	Standard Error (mrem)
1990 (mean = 59.4 mrem, median = 2.0 mrem, SE = 155.9)							
1	Mean (10)	0.9	(0, 389)	0.0	2.50E+04	2.64E+03	167.8
2	Mean (20)	1.5	(0, 389)	0.0	2.48E+04	2.70E+03	167.4
3	Peak (10)	0.1	(0, 388)	0.0	2.51E+04	2.54E+03	167.9
4	Peak (20)	0.4	(0, 389)	0.0	2.52E+04	2.58E+03	168.1
2000 (mean = 4.0 mrem, median = 1.0 mrem, SE = 11.7)							
5	Mean (10)	-0.2	(0, 37)	-0.4	127.5	140.8	17.2
6	Mean (20)	0	(0, 38)	0.0	135.5	135.6	17.3
7	Peak (10)	-0.1	(0, 38)	-0.2	133.1	137.5	17.3
8	Peak (20)	0	(0, 38)	0.0	135.4	134.9	17.2

3.3.3.3. Related Exposures

Two analyses were performed on each of the three shipyards' annual exposure datasets. Using asbestos exposure categories determined by professional judgments, job titles were assigned to either a "High Rank" or "Low Rank" group; missing radiation exposure data from job titles in the Low Rank group were imputed using exposure data from those job titles assigned to the High Rank group. In the first analysis, heavy mobile equipment mechanics were included in the High Rank group. In the second analysis, they were assigned to the Low Rank group. This was done to test the effect of misclassifying a job into an exposure category based on their exposure levels of the proxy exposure.

Estimated Population Mean

The estimates of the population mean exposure were most similar to the true mean exposure for NS3, while the estimates were approximately 10 mrem higher than the true mean for NS1 (Table 3.21). For each of the three yards, there was little difference in population mean bias or confidence interval width when comparing the analysis of the High Rank group with mechanics included to the analysis of the High Rank group without mechanics.

Estimated Population Median

The estimates of the population median exposure were similar to the true median exposure for all six analyses; the estimates were most similar for NS3 (Table 3.21).

Estimated Imputation Variance

The within-imputation variance was considerably higher than the between-imputation variance for all analyses (Table 3.21). The total imputation variance did not appear to differ greatly when comparing the analyses including mechanics in the High Rank group to the analyses excluding mechanics from the High Rank group. The total imputation variance in general overestimated the true variance, with the exception of one analysis for NS2.

Table 3.21. Performance of MI by inclusion status of mechanics (related exposures)

Inclusion status of mechanics in high rank group	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (n=43462, mean = 156.7 mrem, median = 18.0 mrem, SE = 314.2)						
Mechanics included	9.3	(0, 806)	2.0	1.07E+05	1.32	327.0
Mechanics not included	10.5	(0, 810)	3.4	1.08E+05	1.00	328.0
NS2 (n=31532, mean = 222.9 mrem, median = 62.0, SE = 352.1)						
Mechanics included	-2.1	(0, 897)	-0.4	1.19E+05	1.31	345.0
Mechanics not included	2.2	(0, 916)	1.4	1.24E+05	3.17	352.6
NS3 (n=25081, 158.2 mrem, median = 21.0, SE = 315.8)						
Mechanics included	1.8	(0, 789)	0.0	1.03E+05	3.17	321.2
Mechanics not included	0.9	(0, 785)	-0.6	1.02E+05	3.61	319.5

3.3.3.4. Published Literature

Two analyses were performed on each of the three shipyards' annual exposure datasets. Using risk of leukemia as determined in a peer-reviewed publication, job titles were assigned to either a "High Risk" or "Low Risk" group; missing radiation exposure data from job titles in the Low Risk group were imputed using exposure data from those job titles in the High Risk group.

Estimated Population Mean

The estimates of the population mean exposure were similar to the true mean exposure for all three yards; the difference between the estimated and true mean was never greater than 2 mrem (Table 3.22).

Estimated Population Median

The estimates of the population median exposure were similar to the true median exposure for all six analyses (Table 3.22).

Estimated Imputation Variance

The within-imputation variance was considerably higher than the between-imputation variance for all analyses (Table 3.22). The total imputation variance overestimated the true variance for all six analyses.

Table 3.22. Performance of MI using sampling plan based on leukemia risk (published literature)

Shipyard	Mean Bias (mrem)	95% CI of Mean (mrem)	Median Bias (mrem	Within Variance	Between Variance	SE Bias (mrem)
NS1	1.8	(0, 777)	0.8	9.96E+04	0.74	1.5
NS2	1.5	(0, 927)	0.6	1.29E+05	0.64	6.8
NS3	-0.1	(0, 777)	0.0	9.98E+04	3.13	0.2

3.3.4. Alternative Methods for Addressing Missing Data

Twelve of the analyses described above were repeated using both a complete case and mean substitution analysis. The performances of these methods were then compared to the multiple imputation approach.

Complete Case Analysis (CCA)

Table 3.23 summarizes the results of the twelve analyses using complete case analysis. When the data was randomly selected as missing, CCA performed very well in estimating both the mean and median exposure levels. This is to be expected, as complete case analysis is considered an appropriate method when data are truly randomly missing and no sampling biases have occurred. CCA did not perform as well when data were selected to be missing by sample collection date. Using only complete cases in those two analyses meant that the majority of the data were from the recent year group and thus, the mean and median exposure estimates would be underestimates of the true values. When data were evenly sampled by job title, CCA once again performed very well, suggesting that collecting an even number of samples per job title allows for an accurate characterization of the population exposure levels. A bias sampling plan, however, resulted in estimations of the population mean that were 5-15 mrem away from the true mean; the estimated medians were closer to the true medians. In general, the confidence intervals were comparable to what was observed with the multiple imputation approach. The imputation variance calculated for each of the twelve analyses was comparable to the true variance of the datasets.

Mean Substitution

Table 3.24 summarizes the results of the twelve analyses using mean substitution. The estimates of the population mean were very similar to those determined using complete case analysis. However, mean substitution estimated the population median exposure values very poorly. The estimates varied from 15-108 mrem away from the true mean. The confidence intervals were also much narrower in the mean substitution analyses. The imputation variance calculated for each of the analyses was often much lower than the true variance of the datasets.

Comparison to Multiple Imputation

A comparison of the three analytical methods is summarized in Table 3.25. The estimates of the population mean were quite similar between the three methods when the data were missing randomly and when an even percentage of samples were collected from each job title. When the data were missing by sample collection date and by a biased sampling plan base on job title, MI performed the best. The estimates of population median were comparable between MI and CCA; mean substitution performed rather poorly. Finally, the estimated standard error (an expression of the total variance) was comparable between MI and CCA; mean substitution appeared to produce an underestimate of the variance.

Table 3.23. Performance of complete-case analysis

Section ***Sub-analysis***	**Mean Bias (mrem)**	**95% CI of Mean (mrem)**	**Median Bias (mrem)**	**SE Bias (mrem)**
Random *50% missing*	0.1	(0, 172)	0.0	1.6
Random *90% missing*	-0.1	(0, 167)	0.0	-1.0
Random *99% missing*	-1.5	(0, 150)	1.0	-8.7
Random *99.99% missing*	0.8	(0, 140)	0.0	-15.2
Sample Date *Full Analysis*	-22.1	(0, 698)	-6.0	-26.5
Sample Date *HML Analysis*	-22.1	(0, 698)	-6.0	-26.5
Job Title *Even percent, 50%*	0.2	(0, 170)	0.0	0.6
Job Title *Even percent, 5%*	0.0	(0, 177)	0.0	4.1
Job Title *Prior Exposure T1**	9.9	(0, 411)	0.0	18.5
Job Title *Prior Exposure T3**	5.1	(0, 393)	0.0	12.0
Job Title *Related Exposure*	15.0	(0, 827)	4.0	20.5
Job Title *Published Literature*	-4.6	(0, 748)	0.0	-10.1

*Prior Exposure: Trial 1 = top 10% by mean, 1990; Trial 3 = top 10% by peak, 1990

Table 3.24. Performance of mean substitution

Section ***Sub-analysis***	**Mean Bias (mrem)**	**95% CI of Mean (mrem)**	**Median Bias (mrem)**	**SE Bias (mrem)**
Random *50% missing*	0.1	(0, 126)	15.5	-21.8
Random *90% missing*	-0.1	(0, 63)	15.3	-53.9
Random *99% missing*	-1.5	(0, 27)	13.9	-71.4
Random *99.99% missing*	0.9	(15, 17)	16.3	-77.8
Sample Date *Full Analysis*	-22.1	(0, 591)	108.5	-81.2
Sample Date *HML Analysis*	-4.7	(0, 610)	108.5	-80.0
Job Title *Even percent, 50%*	0.2	(0, 125)	15.6	-22.6
Job Title *Even percent, 5%*	0.0	(0, 52)	15.4	-59.9
Job Title *Prior Exposure T1**	9.9	(0, 342)	58	-16.6
Job Title *Prior Exposure T3**	5.1	(0, 344)	23	-13.4
Job Title *Related Exposure*	15	(0, 724)	107	-32.6
Job Title *Published Literature*	-4.6	(0, 665)	77	-50.9

*Prior Exposure: Trial 1 = top 10% by mean, 1990; Trial 3 = top 10% by peak, 1990

Table 3.25. Comparison of the performances of MI, complete-case analysis, and mean substitution

	By MI			***By CCA***			***By Mean Substitution***		
Section ***Sub-analysis***	**Bias of Mean (mrem)**	**Bias of Median (mrem)**	**SE Bias (mrem)**	**Bias of Mean (mrem)**	**Bias of Median (mrem)**	**SE Bias (mrem)**	**Bias of Mean (mrem)**	**Bias of Median (mrem**	**SE Bias (mrem)**
Random *50% missing*	1.2	0.0	4.2	0.1	0.0	1.6	0.1	15.5	-21.8
Random *90% missing*	1.4	0.6	4.3	-0.1	0.0	-1.0	-0.1	15.3	-53.9
Random *99% missing*	0.0	1.0	-6.3	-1.5	1.0	-8.7	-1.5	13.9	-71.4
Random *99.99% missing*	-2.3	0.8	-32.5	0.8	0.0	-15.2	0.9	16.3	-77.8
Sample Date *Full Analysis*	-18.5	-5.2	-21.6	-22.1	-6.0	-26.5	-22.1	108.5	-81.2
Sample Date *HML Analysis*	-0.1	0.4	-4.2	-22.1	-6.0	-26.5	-4.7	108.5	-80.0
Job Title *Even percent, 50%*	1.3	0.0	3.7	0.2	0.0	0.6	0.2	15.6	-22.6
Job Title *Even percent, 5%*	1.4	0.4	6.6	0.0	0.0	4.1	0.0	15.4	-59.9
Job Title *Prior Exposure T1**	0.9	0.0	11.9	9.9	0.0	18.5	9.9	58	-16.6
Job Title *Prior Exposure T3**	0.1	0.0	12.0	5.1	0.0	12.0	5.1	23	-13.4
Job Title *Related Exposure*	9.3	2.0	12.8	15	4.0	20.5	15	107	-32.6
Job Title *Published Literature*	1.8	0.8	1.5	-4.6	0.0	-10.1	-4.6	77	-50.9

*Prior Exposure: Trial 1 = top 10% by mean, 1990; Trial 3 = top 10% by peak, 1990

3.4. Discussion

3.4.1. Random Selection

The first objective of this section was to examine the performance of multiple imputation in estimating population-level exposures when the data are randomly selected as missing. The MI procedure was shown to perform well, even when the percentage of missing data was as high as 99.9% missing. As stated previously, the percentage of exposure measurements missing or uncollected in an occupational study is likely to be high; thus, these analyses reflect plausible scenarios. In addition, using multiple imputation to estimate exposure levels using data with increasing percentages of missing data is a good exercise in understanding the overall performance abilities and limitations of MI. While there does appear to be a point at which the percentage of missing data may be too high for the MI estimates to be considered reliable, the analyses in this section suggest that population-level exposure levels may be reasonably estimated when only a fraction of the work population has been randomly sampled. This observation may benefit both industrial hygienists in the field as well as research epidemiologists. If a researcher is faced with limited exposure data and has reason to believe those data are available at random, an approach such as multiple imputation may be appropriate for use in developing accurate population exposure estimates.

The positive performance of MI was observed across all three yards, reinforcing its potential as a viable approach. Finally, the total imputation variance was shown to be similar to the true variance in the datasets. This confirms a widely noted strength of the multiple imputation procedure and makes MI a stronger option when working with missing exposure data.

3.4.2. Selection by Sample Collection Date

The first objective of this section was to examine how the accuracy of population-level exposure estimates is affected by the use of exposure data from multiple time periods and how this impacts the performance of a multiple imputation approach. When stratifying the data into two time periods based on sample collection date, it was observed that the more recent exposures were lower in exposure levels than the earlier ones. Thus, when using the recent year data to impute the missing early year data, the estimated population exposure levels were underestimates of the true exposure levels.

Such a scenario is likely to be common in occupational cohorts; exposure levels do tend to decrease with time as worker protection options improve. This requires the researcher to consider possible trends in exposure levels over time and whether the available exposure data accurately represents the expected exposure levels of missing or uncollected data. Given the expected changes in exposure levels over time, multiple imputation must be used carefully; a subset of the available exposure measurements may be more appropriate.

Thus, the second objective of this section was to compare the performance of MI when three varying subsets of exposure data were available for use for the imputations. Of the three, the Bin Analysis performed the best across all three shipyards. This is not surprising, as an attempt was made to reflect the exposure levels of the early year data. By using a subset of the recent year data that mirrored the proportions of early year exposure measurements placed in each bin, the imputed missing data more accurately reflected the true missing

exposure levels. This then resulted in population-level estimates that more closely resembled the true levels.

Correctly assigning recent year exposure data to each of the six bins is not a challenge when the artificially "missing" exposure data are in fact actually available, but in real-world scenarios where the values of the missing data are truly unknown, such a task would be nearly impossible. To address this, the HML analysis was performed, in which recent year exposures were assigned to one of only three bins based on a qualitative exposure scale of High, Medium, and Low. Even without knowing the quantitative values of the missing exposure data, it may be possible to predict whether a value would fall into the High, Medium, or Low category based on information available through other variables and using professional judgment. Such a technique is quite common in occupational studies in which quantitative measurement data are unavailable (Ramachandran et al. 2003). The HML analysis performed nearly as well as the Bin analysis and would likely be considered a much more feasible approach.

Finally, the Non-zero analysis also performed better than simply using all of the available recent year data. Again, this is not surprising, as the high percentage of zero values in the recent year data was the reason these exposure levels were much lower than the early year exposure levels. This exercise was performed to illustrate how easily MI is influenced by the exposure levels of the data used in the imputation process and to highlight the importance of carefully examining the data prior to performing any MI analysis. However, it is not

practical, nor wise, to simply removed a large proportion of the exposure data and this particular analysis is not recommended.

The analyses performed in this section provide some options to epidemiologists and exposure assessors working with sparse historical measurement data. As discussed in the beginning of this section, researchers may often find themselves working with exposure data from a different time period than the one of interest to the study. While this data can still prove useful in estimating population exposures, and MI is still a viable option to address the missing data problem, some careful consideration of the possible differences in exposure levels between the available and missing data is necessary.

3.4.3. Selection by Job Title

The objective of this section was to assess the ability of limited sampling by job title to accurately characterize the exposure profile of the overall work population and to examine the potential for multiple imputation to assist in characterizing such population exposures. To do this, various plausible industrial hygiene sampling plans were explored. In the first set of analyses, an even percentage of samples were collected from each job title in the work population. Although the percentage of samples collected from each job was decreased to just 1% per job, the MI estimates were still quite accurate, suggesting that collecting evenly from each job title may be an appropriate approach when the number of allotted samples is small.

In the remaining analyses, sampling plans were designed around a sampling strategy in which intentional oversampling from certain job titles based on some *a priori* knowledge of exposure occurred. These types of sampling plans are commonly utilized in the field; industrial hygienists want to sample in the most effective and efficient way, given the number of samples they can feasibly collect. Of the three sampling plans investigated, sampling based on prior exposure levels performed the best when approached with an MI method. This seems logical, as quantitative exposure data from the same work site and/or work population is always the preferred source of exposure data, when available. Although exposure levels do often change over time, these analyses suggest that long-term patterns of exposure levels by job title remain fairly constant. Sampling plans were designed for both the year 1990 and 2000 based on mean and peak exposure levels; many of the same job titles that were identified as having the highest exposure levels from 1979-1989 were also found to have the highest exposures from 1989-1999.

When prior exposure data are not available, other sources of exposure information can be helpful, even if they are less reliable. Asbestos exposures, which may be considered related to radiation exposures, were used to help design a set of sampling plans. Those jobs that were identified as having high asbestos exposures, based on professional judgments, were assumed to have high radiation exposures as well. This strategy makes two big assumptions: that the professional judgments are reliable, and that high asbestos exposures correlate to high radiation exposures. There have been a number of research studies that have looked at the reliability of professional judgments (Hawkins & Evans, 1989; Seel et al. 2007; Vadali et al. 2009). This approach is dependent on the experts' level of familiarity with the relevant jobs,

tasks, and exposures. The rationale behind judgments can sometimes be unclear and it is often impossible to comment on the validity of these subjective estimates, as there is seldom any measurement data, or "gold standard," for comparison. Nevertheless, this is sometimes the only source of exposure information available and thus it is worth investigating through a multiple imputation technique.

The other assumption, that high asbestos exposures correlate to high radiation exposures, was explored within the analysis. The jobs that were identified as having the highest asbestos exposures did not necessarily have the highest radiation levels (for example, heavy mobile equipment mechanics). However, the MI estimates were still fairly close to the true estimates and an analysis in which heavy mobile equipment mechanics were removed from the high exposure group did not significantly change the results. This suggests that such an approach may be robust to a few misclassified jobs.

The final analysis used cancer risks published in the peer-reviewed literature to help inform the sampling plan. Similar to the asbestos exposure analysis, this strategy makes the assumption that those jobs with increased cancer risk are also the jobs with high radiation exposures. Again, however, the MI estimates were close to the true values, suggesting this might also be a viable approach when quantitative measurement data are unavailable.

The analyses conducted in this section all acknowledge that industrial hygienists are often faced with designing sampling plans with a limit on the number of samples to be collected and that the workers' job titles are often used as the deciding factor in whether to sample.

Based on the results of this section, there appear to be a number of options that may be appropriate for the design of future sampling plans.

3.4.4. Alternative Methods for Addressing Missing Data

Complete case analysis, while not generally a preferred method, may be appropriate when the data are truly missing at random and the complete cases are therefore a random sample of all cases. As shown in the analyses in which the data were selected to be randomly missing, complete case analysis and the MI approach perform comparably in such a scenario. However, for most occupational cohorts, exposure data will not be missing completely at random and thus CCA is no longer an appropriate choice. This is illustrated in the analyses in which exposure data were missing not randomly but by sample collection date or job title. In these cases, the MI approach performed much better than CCA. Given that most sampling plans likely have some sampling bias associated with them, MI should generally be considered a superior option over CCA.

Mean substitution, while considered an improvement over CCA, still has a number of disadvantages as compared to MI. One of the most concerning, which was illustrated in these analyses, is the underestimation of the variance of the data. Multiple imputation has a number of advantages, as discussed in Chapter 1, and maintaining the variability of the data is a significant one. Another drawback to mean substitution is how greatly it affects estimates of the median exposure levels, since all missing data are replaced with one single value, the sample mean. Thus, like CCA, mean substitution should generally be considered inferior to multiple imputation.

3.5. Conclusion

The results of the analyses in Chapter 3 suggest that an MI approach can perform well when data are missing randomly, even when the percentage of missing data is high. These analyses also highlighted a major advantage of multiple imputation – that is, that the total imputation variance is similar to the true variance of the data. In addition, a section of the analyses were performed to illustrate the impact changes exposure levels over time can have on a model-based method such as multiple imputation. An important conclusion from these analyses is that it is important to be aware of potential differences between the work population with available and missing data in order to properly characterize the exposure levels of the population. Finally, when working with occupational exposure data, the missing data patterns may be based on the perceived exposure levels of each job title, meaning the available data may be biased in some way that will be reflected in the estimated exposures. The different sampling strategies simulated in this chapter all performed reasonably well. This suggests that there are likely a number of appropriate sampling designs and that even when available data are biased by job title, multiple imputation is a viable option.

CHAPTER 4: CHARACTERIZING THE EXPOSURE PROFILE OF AN SEG

4.1. Introduction

4.1.1. Background

4.1.1.1. Similar Exposure Groups

In many occupational epidemiology studies, individual exposure data are limited or even non-existent. Some of the more common reasons for this include financial constraints permitting only a limited number of samples to be collected; restricted availability of the population to be monitored, sometimes due to changing work shifts or tasks completed in confined areas; or reliance on historical data, which can often be sparse. In addition, occupational exposure measurements are often originally collected not for research purposes but for compliance determinations, which may rely on only a few samples per worker when characterizing exposure.

When industrial hygienists are faced with defining exposures from only a limited number of samples, they will often attempt to estimate exposure levels using grouped data. One of the most common approaches, particularly for occupational cohorts, is to divide the population into Similar Exposure Groups (SEGs), or clusters of workers believed to have the same general exposure profile for the agent(s) under study, from which individual exposure levels are then established. This approach pools available measurement data across individuals within each group to create grouped estimates; each worker's exposure level is then determined using the measurement data available for his/her relevant SEG(s) (Werner &

Attfield, 2000). These exposure groups are also sometimes referred to as homogeneous exposure groups (HEGs).

4.1.1.2. Definition and Calculation

SEGs are typically defined using industrial hygienists' observations of the work process, job, task, and/or environmental agent(s). The American Industrial Hygiene Association (AIHA) defines SEGs as "groups of workers having the same general exposure profile for the agent(s) being studied because of the similarity and frequency of the tasks that they perform, the materials and processes with which they work, and the similarity of the way that they perform the tasks" (Dinardi, 2003).

Some attempt has been made to construct a quantitative definition of SEGs. Rappaport defined a uniformly exposed group as a group of workers in which

$$\frac{\bar{x}_{97.5\%tile}}{\bar{x}_{2.5\%tile}} \leq 2$$

where the values are the 97.5th and 2.5th percentiles of the distribution of the individual worker means (Rappaport, 1991). This ratio creates a quantitative limit on the acceptable variability of the distribution of individual exposure means within an exposure group. However, it has been noted that small sample sizes, lognormality, and classification based solely on worker job description can all lead to large variation in values of the ratio, highlighting the challenges hygienists are faced with when creating SEGs (Perkins, 1997).

4.1.1.3. Strengths and Limitations

Classifying workers into SEGs is a valuable time-saving measure for industrial hygienists and often a financial necessity. In addition, compared to a common alternative approach of sampling only the assumed highest risk populations, creating SEGs is a theoretically favorable method as it allows for all of the available measurement data to be used, attempts to account for differences in work tasks and practices, and potentially provides exposure information for the entire worker population, including those with low and intermediate exposures (Corn & Esmen, 1979).

However, in using this method, industrial hygienists make the assumption that exposures within one group are statistically similar enough that, by collecting measurements for a small number of individuals in the group, the exposures of the remaining workers can be defined (Loomis & Kromhout, 2004). Classifying workers into exposure groups implicitly assumes that the probability and distribution of exposures is uniform for all members of the group. Should this assumption of homogeneity not be true, there is a risk of misclassifying workers' true exposure levels.

Several studies examining the homogeneity of SEGs have shown that many occupational groups are not as uniformly exposed as was generally assumed by the hygienists (Burdorf & van Tongeren, 2003 ; Kromhout et al. 1993; Rappaport et al. 1993). This is likely because similarities in observational factors such as work environment and job description are generally not sufficient to assign workers to the correct homogeneous exposure groups without the availability and consideration of quantitative exposure data (Stewart & Stenzel,

2001). Exposure variability within a theoretical SEG can also result from sources often overlooked, including inconsistent work practices, day-to-day random variability, and within- and between-individual variability.

4.1.1.4. SEGs in Epidemiology

Another concern when relying on SEGs for exposure estimation comes from the intended purpose of the exposure groupings. Industrial hygienists most often design sampling plans and collect measurements in order to determine compliance with regulatory standards. Thus, they may collect limited information on the worker. Should this exposure data later be used in an epidemiology study, the lack of additional information on the worker population now becomes a detriment. Indeed, exposure groups defined for epidemiology studies are often identified based on categories of information chosen for practical reasons – such as availability of information – rather than their appropriateness, which may result in groupings that are not homogeneous (Stewart & Stenzel, 2001). The disconnect between the industrial hygiene sampling plan and the research aims of an occupational exposure study make using SEGs a potential source of exposure misclassification. Better communication between the hygienists and the investigators regarding the necessary information to be collected may be one way to improve the effectiveness of SEGs in exposure estimation.

4.1.2. Specific Aims

Despite the concerns discussed above, the SEG strategy will no doubt continue to be used in the field and thus an attempt should be made to better understand its capabilities. Thus, the specific aims of the chapter are to examine how SEGs are affected by various sampling

plans, explore additional workplace variables that may influence the homogeneity of an SEG, and test the performance of a multiple imputation approach in estimating SEG-level exposures.

4.1.3. Focused Research Objectives

The analyses performed in Chapter 4 have been summarized in Table 4.1. In order to test the performance of the multiple imputation method, artificially missing exposure data were generated from each complete dataset. The method by which data were selected to be missing was varied for each analysis, with the goal being to generate missing data in ways that reflect real-world sampling scenarios. The overall objectives and challenges faced by industrial hygienists and epidemiologists when collecting and/or reviewing exposure data were considered. Data were ultimately assigned to be missing based on one of three general selection patterns. For the first set of analyses, data were randomly selected to be missing within a given time period or time interval. In the second set of analyses, data were selected to be missing to achieve a desired percentage of sampled workers. In the final set of analyses, data were selected as missing based on the value of the model covariates *worker birth year* and *sample collection quarter*. Each of the three selection patterns is described in detail in the following sections. Unless otherwise stated, analyses were performed on the data from all three shipyards.

4.1.3.1. Grouping Measurements by Time Intervals

In the first section, data are grouped into time intervals based on the sample collection year. The analyses in this section address the following research objectives:

- Examine the ability of using a limited number of measurements per year to accurately characterize the exposure profile of a SEG through a multiple imputation approach
- Explore the potential of grouping measurements into larger time intervals to assist in estimating exposures within an SEG

4.1.3.2. Variation in Number of Samples Collected

In the second section, the total number of samples collected within an SEG is varied by the number of workers sampled and number of samples collected per worker. The analyses in this section address the following research objective:

- Understand the influence various sampling strategies have on the ability to accurately estimate the exposure profile of an SEG using multiple imputation

4.1.3.3. Exploring Additional Exposure Covariates

In the third and final section, additional information on the work population, beyond job title, will be considered in defining SEGs. The analyses in this section address the following research objectives:

- Explore whether the variables *worker birth year* and *sample collection quarter* can offer additional information on the exposure levels of workers within an SEG
- Comment on the true homogeneity of exposures within an assumed SEG

4.1.4. Study Population

The study population consisted of shipyard workers from three naval shipyards who held at least one of the three selected job titles: pipefitting, welding, and electrician. For these analyses, a similar exposure group (SEG) is defined as all workers holding the specified job title. Workers holding the job titles of interest were identified from the larger study population described in Chapter 2; thus, the same selection criteria were applied.

A summary of the number of radiation measurements collected within each job title is provided in Table 4.2. The difference between the daily and annual radiation measurements was detailed in Chapter 2. Tables 4.3-4.5 summarize the exposure levels observed within each job title for each of the three yards.

Table 4.1 Summary of analyses completed in Chapter 4

Section	Missing Data Pattern	Sub-analysis	Exposure Data Used
4.1 Grouping Measurements by Time Intervals	By random	*Sample Collection Year* variable: each year	Daily
		Sample Collection Year variable: 5-year time interval	Daily
		Sample Collection Year variable: 10-year time interval	Daily
4.2 Variation in Number of Samples Collected	By random within a set % of workers sampled	*Pipefitters* – from 5-100% workers sampled	Daily
		Welders – from 5-100% workers sampled	Daily
4.3 Additional Exposure Covariates	By random	Effect of removing variables from model: *birth year, education, race*	Annual
		Effect of including variables in model: *sample collection year, quarter, and both*	Daily

Table 4.2. Number of daily and annual measurements by SEG

	Daily Measurements			Annual Measurements		
Job Title (SEG)	Naval Shipyard #1 (NS1)	Naval Shipyard #2 (NS2)	Naval Shipyard #3 (NS3)	Naval Shipyard #1 (NS1)	Naval Shipyard #2 (NS2)	Naval Shipyard #3 (NS3)
Pipefitting						
Number of measurements for analyses	72337	49716	44959	7108	4466	3448
Size of work population for analyses*	999	742	420	999	737	420
Welding						
Number of measurements for analyses	24646	23831	18154	2387	1709	1287
Size of work population for analyses*	355	255	161	352	247	158
Electrician						
Number of measurements for analyses	37065	24501	19837	4291	3124	2273
Size of work population for analyses*	752	563	338	751	562	338

*The difference in size of work population between daily and annual datasets is due to the removal from the annual dataset of workers who switched jobs mid-year

Table 4.3. Daily and annual exposure measurements: pipefitting SEG

Shipyard	NS1	NS2	NS3
Daily Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	35101	21399	27813
% of 0 mrem values	48.5	43.0	61.9
Mean (mrem)	12.5	18.4	13.2
Median (mrem)	1.0	1.0	0.0
Peak (mrem)	1782	2238	2960
Annual Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	1439	549	692
% of 0 mrem values	20.2	12.3	20.0
Mean (mrem)	130.0	203.1	172.0
Median (mrem)	24.0	78.0	41.0
Peak (mrem)	1782	2259	2960

Table 4.4. Daily and annual exposure measurements: welding SEG

Shipyard	NS1	NS2	NS3
Daily Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	12048	8866	10359
% of 0 mrem values	48.9	37.2	57.1
Mean (mrem)	16.4	19.9	12.3
Median (mrem)	1.0	1.0	0.0
Peak (mrem)	1505	1856	2485
Annual Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	431	178	193
% of 0 mrem values	18.1	10.4	15.0
Mean (mrem)	167.3	194.0	172.7
Median (mrem)	40.0	46.0	102.0
Peak (mrem)	1764	1964	2485

Table 4.5 Daily and annual exposure measurements: electrician SEG

Shipyard	NS1	NS2	NS3
Daily Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	19355	12077	13092
% of 0 mrem values	52.2	49.3	66.0
Mean (mrem)	10.3	24.3	12.7
Median (mrem)	0.0	1.0	0.0
Peak (mrem)	1770	2456	2839
Annual Measurements			
Start Year	1975	1975	1976
End Year	2005	2005	2005
Number of 0 mrem values	1235	513	681
% of 0 mrem values	28.8	16.4	30.0
Mean (mrem)	90.5	188.6	110.9
Median (mrem)	7.0	40.0	12.0
Peak (mrem)	1770	2456	2839

4.2. Methods

4.2.1. Grouping Measurements by Time Intervals

The number of samples collected per year within an assumed SEG is often limited. Whether an industrial hygienist is trying to determine compliance or an epidemiologist is trying to calculate risk estimates, there may be a point at which the number of samples collected per year within SEG becomes too few to develop accurate estimates, even if those samples are randomly collected. In addition, the performance of a modeling approach, such as the multiple imputation method utilized throughout this dissertation, may suffer. One possible solution may be to group measurements into broader time intervals, creating larger sample sizes per interval, prior to analysis.

In the first section, data are grouped into time intervals based on the sample collection year. The analyses in this section address the following research objectives:

- Examine the ability of using a limited number of measurements per year to accurately characterize the exposure profile of a SEG through a multiple imputation approach
- Explore the potential of grouping measurements into larger time intervals to assist in estimating exposures within an SEG

One of the variables used in the imputation models described throughout this dissertation is the sample collection year. When estimating exposures within an SEG, the number of measurements available for any given year may be small, even before missing data are simulated. The second column in Table 4.6 (*0% missing*) displays the true number of daily radiation measurements that were collected each year on pipefitters at NS1. As the proportion of randomly selected artificial missing exposure data is increased, the number of available measurements per year decreases. When 99.95% of the exposure data are missing, for example, many of the years contain only one available measurement. Such a small sample size per year may have an effect on the overall ability of the multiple imputation approach to accurately estimate the exposure levels of the SEG over time.

A possible solution may be to group the available exposure measurement data into larger time intervals. In Table 4.7, the same daily radiation measurements collected on pipefitters at NS1 have now been grouped together based on 5-year time intervals. In Table 4.8, these measurements have now been grouped together based on 10-year time intervals. Even as the

proportion of randomly selected missing exposure data is increased, the sample size per interval remains relatively large.

There are now three different ways to categorize the sample collection date variable: by year; by 5-year intervals; and by 10-year intervals. The analyses in this section will compare the performance of the multiple imputation approach when each of the three sample collection date variables is used and the percent of missing data is varied between 50% and 99.95%. In each scenario, the remaining available data will be used to impute the missing values. Although not shown, the proportion of exposure measurements collected per year on pipefitters at NS2 and NS3 looked similar. Thus, these analyses were performed on pipefitter exposure data from all three yards.

Table 4.6. Number of daily measurements available per year at NS1 by percentage of missing data: pipefitting SEG

	Number of available measurements						
Year	*0%*	*50.0%*	*90.0%*	*95.0%*	*99.0%*	*99.90%*	*99.95%*
1975	2	1	1	1	1	1	1
1976	74	41	3	2	1	1	1
1977	67	41	3	1	1	1	1
1978	59	38	8	2	1	1	1
1979	60	36	4	2	1	1	1
1980	49	25	4	2	1	1	1
1981	48	24	10	4	1	1	1
1982	345	154	40	13	3	1	1
1983	354	179	32	21	7	2	2
1984	1146	570	114	58	9	1	1
1985	980	472	110	48	11	1	1
1986	992	514	93	42	15	1	1
1987	923	449	100	48	8	1	1
1988	1034	518	112	61	7	1	1
1989	1060	515	109	60	18	1	1
1990	1200	594	123	68	9	1	1
1991	1288	649	128	69	13	2	1
1992	1538	797	154	71	14	1	1
1993	1316	660	125	63	8	1	1
1994	864	430	83	41	12	3	2
1995	760	371	64	35	10	2	1
1996	941	475	99	59	11	1	1
1997	708	369	68	34	5	1	1
1998	617	306	58	38	9	1	1
1999	11635	5839	1174	580	112	14	5
2000	6400	3203	603	322	75	10	6
2001	6693	3315	697	348	71	4	1
2002	7639	3824	745	360	82	12	5
2003	7460	3689	776	385	80	12	6
2004	8106	4003	836	405	83	8	3
2005	7979	4006	797	398	85	5	2
Total	**72337**	**36107**	**7273**	**3641**	**764**	**94**	**54**

Table 4.7. Number of daily measurements available per 5-year interval at NS1 by percentage of missing data: pipefitting SEG

	Number of available measurements						
Year Interval	*0%*	*50.0%*	*90.0%*	*95.0%*	*99.0%*	*99.90%*	*99.95%*
1975-1980	311	182	22	9	3	1	1
1981-1985	2873	1399	306	144	30	4	4
1986-1990	5209	2590	537	279	57	3	2
1991-1995	5766	2907	554	279	57	7	4
1996-2000	20301	10192	2002	1033	212	26	13
2001-2005	37877	18837	3851	1896	401	41	16
Total	**72337**	**36107**	**7272**	**3640**	**760**	**82**	**40**

Table 4.8. Number of daily measurements available per 10-year interval at NS1 by percentage of missing data: pipefitting SEG

	Number of available measurements						
Interval	*0%*	*50.0%*	*90.0%*	*95.0%*	*99.0%*	*99.90%*	*99.95%*
1975-1985	3184	1635	331	164	32	3	2
1986-1995	10975	5454	1100	515	102	13	8
1996-2005	58178	29018	5841	2961	626	65	29
Total	**72337**	**36107**	**7272**	**3640**	**760**	**81**	**39**

4.2.2. Variation in Number of Samples Collected

Industrial hygienists utilize similar exposure groups as a way to estimate the exposure levels for a large group of workers when only a limited number of samples can be collected. Thus, when designing a sampling plan for within an SEG, the industrial hygienist has to make several decisions, including how many samples to collect and from how many workers, in order to be able to accurately assess exposure. In most cases, sampling from 100% of the SEG population is not practical. Instead, the hygienist must decide what percentage of the population they can realistically sample and how many samples should be collected per worker. This section aims to understand how different sampling plans can affect the estimations of exposure within an SEG and to investigate whether multiple imputation is an appropriate approach to assist in developing such estimates.

In this section, the total number of samples collected within an SEG is varied by the number of workers sampled and number of samples collected per worker. The analyses in this section address the following research objective:

- Understand the influence various sampling strategies have on the ability to accurately estimate the exposure profile of an SEG using multiple imputation

For these analyses, various hypothesized industrial hygiene sampling plans were created for two separate SEGs: pipefitting and welding. The daily radiation measurements collected in 1990 for each SEG will be used. A summary of the true exposure levels for each SEG in 1990 is summarized in Table 4.9.

The 14 different sampling plans designed for these analyses are summarized for pipefitters in Table 4.10 and for welders in Table 4.11. For both SEGs, the percentage of workers classified as "sampled" is varied from 5-100% and the number of samples collected per worker is varied from 1-4 samples. The percentage of total measurements assigned as "sampled" therefore varied from 1.0% to 63.0% for pipefitters and 1.0% to 65.0% for welders. Those measurements that were not designated as "sampled" in the sampling plans were considered "not sampled."

Two separate exercises were examined in this section; each exercise considered all 14 sampling plans. In the first exercise, only those measurements that were assigned as "sampled" were used to calculate the exposure estimates for that SEG. This strategy reflects how industrial hygienists commonly assess exposure in the field. The collected samples are

assumed to be representative of all exposures in that SEG, including those of workers on whom measurements were not collected. In the second scenario, those same measurements that were assigned as "not sampled" were imputed. The estimated exposure levels determined by imputation were then compared to those estimated in the first exercise.

Table 4.9. Summary of daily measurements in 1990 by SEG

Shipyard	Total No. of measurements collected in 1990	Total No. of workers in 1990	Mean exposure (mrem)	Median Exposure (mrem)
Pipefitting				
NS1	1200	377	43.1	3.0
NS2	761	221	53.2	22.0
NS3	607	160	31.0	3.0
Welding				
NS1	351	105	76.1	17.0
NS2	240	69	89.4	54.0
NS3	166	54	28.8	9.5

Table 4.10. Summary of the 14 sampling plans designed for the pipefitting SEG

Shipyard	NS1 (n=1200 meas.)		NS2 (n=761 meas.)		NS3 (n=607 meas.)	
% workers sampled (No. samples per worker)	*No. workers sampled*	*No. meas. collected (% of total)*	*No. workers sampled*	*No. meas. collected (% of total)*	*No. workers sampled*	*No. meas. collected (% of total)*
100% (1)	377	377 (31.4)	221	221 (29.0)	160	160 (26.4)
100% (2)	377	684* (57.0)	221	412* (54.1)	160	307* (50.6)
50% (1)	189	189 (15.8)	111	111 (14.6)	80	80 (13.2)
50% (2)	189	378 (31.5)	111	222 (29.2)	80	160 (26.4)
50% (4)	189	756 (63.0)	111	444 (58.3)	80	320 (52.7)
20%(1)	75	75 (6.3)	44	44 (5.8)	32	32 (5.3)
20% (2)	75	150 (12.5)	44	88 (11.6)	32	64 (10.5)
20% (4)	75	300 (25.0)	44	176 (23.1)	32	128 (21.1)
10% (1)	38	38 (3.2)	22	22 (2.9)	16	16 (2.6)
10% (2)	38	76 (6.3)	22	44 (5.8)	16	32 (5.3)
10% (4)	38	152 (12.7)	22	88 (11.6)	16	64 (10.5)
5% (1)	19	19 (1.6)	11	11 (1.4)	8	8 (1.3)
5% (2)	19	38 (3.2)	11	22 (2.9)	8	16 (2.6)
5% (4)	19	76 (6.3)	11	44 (5.8)	8	32 (5.3)

*Some workers only had 1 measurement available; thus, the total number of samples collected is less than 2x work population

Table 4.11. Summary of the 14 sampling plans designed for the welding SEG

Shipyard	NS1 (n=351 meas.)		NS2 (n=240 meas.)		NS3 (n=166 meas.)	
% workers sampled (No. samples per worker)	*No. workers sampled*	*No. meas. collected (% of total)*	*No. workers sampled*	*No. meas. collected (% of total)*	*No. workers sampled*	*No. meas. collected (% of total)*
100% (1)	105	105 (29.9)	69	69 (28.8)	54	54 (32.5)
100% (2)	105	194* (55.3)	69	131* (54.6)	54	98* (59.0)
50% (1)	53	53 (15.1)	35	35 (14.6)	27	27 (16.3)
50% (2)	53	106 (30.2)	35	70 (29.2)	27	54 (32.5)
50% (4)	53	212 (60.4)	35	140 (58.3)	27	108 (65.1)
20%(1)	21	21 (6.0)	14	14 (5.8)	11	11 (6.6)
20% (2)	21	42 (12.0)	14	28 (11.7)	11	22 (13.3)
20% (4)	21	84 (23.9)	14	56 (23.3)	11	44 (26.5)
10% (1)	11	11 (3.1)	7	7 (2.9)	6	6 (3.6)
10% (2)	11	22 (6.3)	7	14 (5.8)	6	12 (7.2)
10% (4)	11	44 (12.5)	7	28 (11.7)	6	24 (14.5)
5% (1)	5	5 (1.4)	†	†	†	†
5% (2)	5	10 (2.8)	†	†	†	†
5% (4)	5	20 (5.7)	†	†	†	†

*Some workers only had 1 measurement available; thus, the total number of samples collected is less than 2 x work population

† A 5% sample was not collected as it would have resulted in too small of a sample size for analysis.

4.2.3. Exploring Additional Exposure Covariates

As discussed in the introduction, a similar exposure group may be defined differently based on the purpose of the exposure data. In many cases, information is not available on all variables potentially related to exposure. When occupational exposure data are used for an epidemiology study, SEGs are generally created based on available information, which may be limited beyond job title and exposure level. However, it is possible that additional information on the exposures levels of the work population exists within variables readily available but not commonly considered. If identified, these variables may assist in developing improved estimates of the exposure profile of the study population. This section aims to explore whether two specific variables – *worker birth year* and *sample collection quarter* – can potentially offer valuable insights into the variability of exposures within an SEG.

In the third and final section, additional information on the work population, beyond job title, will be considered in defining SEGs. The analyses in this section address the following research objectives:

- Explore whether the variables *worker birth year* and *sample collection quarter* can offer additional information on the exposure levels of workers within an SEG
- Comment on the true homogeneity of exposures within an assumed SEG

Birth Year

When developing exposure models within an SEG for the shipyard population, the following variables have been considered: sample collection year, sample collection quarter, worker birth year, worker educational level, and worker race. Additional information for each employee, such as work location or specific work task, is desirable but unavailable. As mentioned in a previous section, SEGs for this population are defined based on job title, largely due to limited availability of additional information regarding the worker's task. However, in this section, the variable *birth year* is explored as a potential surrogate for work task.

Table 4.12 stratifies the 1990 annual radiation measurements for pipefitters in all three yards by birth year. Bin 1 contains all pipefitters who, in 1990, had a birth year prior to 1955; Bin 2 contains all pipefitters who had a birth year of 1955 or later. In all three yards, those pipefitters who were born prior to 1955 had lower mean and peak exposure levels; at NS1 and NS2, the median exposure levels were also lower. The same pattern was observed when the daily radiation measurements were examined (Table 4.13). To investigate this pattern further, the same pipefitter population was stratified into six bins by birth year (Table 4.14). As the birth year of the pipefitter became more recent, the mean exposure levels generally increased. The same patterns were also observed when welders (Tables 4.15-4.17) and electricians (Tables 4.18-4.20) were examined. Thus, it appears that in 1990, the younger workers received higher mean and peak exposure levels of radiation compared to the older workers. The hypothesis is that the younger skilled trades workers are more likely to be apprentices or journeymen in their craft and perform more hands-on work. The older skilled

trades worker are more likely to be masters, or supervisors, and spend more time managing, observing, and offering guidance.

It is therefore possible that the variable *birth year* may be a surrogate for work task and could thus be an important variable to collect, particularly when detailed information on the workers' job duties is not available. To test the effect *birth year* has on the imputation exposure models, four different models were fit using the annual radiation measurements available for each SEG. In the first model, the variables *sample collection year*, *birth year, education level*, and *race* were all used to impute the randomly selected 50% missing data. In the second model, *birth year* was removed. In the third model, *education level* was removed; in the fourth model, *race* was removed. These analyses were performed using the pipefitting SEG.

Sample Collection Quarter

When working with the daily radiation measurements, the sample collection date can be described by both year and quarter. While changes in exposure levels over a period of years is to be expected, changes within a year (per quarter) may be subtler. However, such fluctuations could offer insights into the overhaul and maintenance schedule of the yards' submarines, information that is not currently available but would further help to describe the employees' work tasks. Changes per quarter may also highlight seasonal changes in exposure levels. To test the effect quarter has on the imputation exposure models, three different models were fit using the daily radiation measurements available for each SEG. In the first model, the variables *sample collection year* and *sample collection quarter* were both used as

time variables to impute a randomly selected 50% missing data. In the second model, only *sample year* was included; in the third model, only *sample collection quarter* was included. In each model, the variables *birth year, education level,* and *race* were also added. These analyses were performed using the pipefitting SEG.

Table 4.12. Annual measurements stratified by birth year: pipefitting SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: Prior to 1955	174	106.0	15.0	1553
Bin 2: 1955 or later	188	176.6	53.0	1604
NS2				
Bin 1: Prior to 1955	110	152.0	46.0	1404
Bin 2: 1955 or later	100	236.2	208.5	1455
NS3				
Bin 1: Prior to 1950	64	111.4	40.0	539
Bin 2: 1950 or later	89	130.8	29.0	552

Table 4.13. Daily measurements stratified by birth year: pipefitting SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: Prior to 1955	572	32.4	2.0	1255
Bin 2: 1955 or later	628	52.9	6.0	1322
NS2				
Bin 1: Prior to 1955	395	42.7	13.0	1044
Bin 2: 1955 or later	366	64.6	32.0	1049
NS3				
Bin 1: Prior to 1950	260	27.6	3.0	467
Bin 2: 1950 or later	347	33.6	3.0	453

Table 4.14. Annual measurements stratified into six bins by birth year: pipefitting SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
		NS1		
Bin 1: 1932-1937	7	18.7	9.0	63
Bin 2: 1938-1943	17	21.1	3.0	100
Bin 3: 1944-1949	66	114.6	26.0	697
Bin 4: 1950-1955	98	118.1	11.0	1553
Bin 5: 1956-1961	112	162.3	41.0	1571
Bin 6: 1962-1969	62	223.2	131.0	1604
		NS2		
Bin 1: 1928-1938	15	93.3	47.0	275
Bin 2: 1939-1944	20	125.5	35.0	642
Bin 3: 1945-1950	46	174.5	44.0	1404
Bin 4: 1951-1956	45	210.3	110.0	1316
Bin 5: 1957-1962	70	237.7	191.5	1455
Bin 6: 1963-1968	14	164.4	212.0	333
		NS3		
Bin 1: 1933-1938	9	72.7	16.0	306
Bin 2: 1939-1944	18	63.1	21.5	231
Bin 3: 1945-1950	49	125.6	54.0	539
Bin 4: 1951-1956	60	134.2	33.5	455
Bin 5: 1957-1960*	17	163.5	28.0	552

*For NS3, only five bins were created

Table 4.15. Annual measurements stratified by birth year: welding SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
		NS1		
Bin 1: Prior to 1955	40	112.8	64.5	836
Bin 2: 1955 or later	58	378.5	152.0	1654
		NS2		
Bin 1: Prior to 1955	33	236.5	264.0	552
Bin 2: 1955 or later	34	400.0	439.0	600
		NS3		
Bin 1: Prior to 1953	27	77.1	77.0	214
Bin 2: 1953 or later	26	103.6	89.5	328

Table 4.16. Daily measurements stratified by birth year: welding SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: Prior to 1955	130	36.4	7.5	705
Bin 2: 1955 or later	221	99.5	26.0	1252
NS2				
Bin 1: Prior to 1955	107	73.1	40.0	310
Bin 2: 1955 or later	133	102.6	68.0	430
NS3				
Bin 1: Prior to 1953	89	23.4	9.0	182
Bin 2: 1953 or later	77	35.0	13.0	197

Table 4.17. Annual measurements stratified into six bins by birth year: welding SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: 1937-1945	5	32.4	5.0	133
Bin 2: 1946-1951	19	131.6	41.0	836
Bin 3: 1952-1957	28	185.0	120.0	756
Bin 4: 1958-1963	32	346.7	160.5	1521
Bin 5: 1964-1969	14	537.6	148.0	1654
NS2				
Bin 1: 1928-1942	6	339.2	403.5	552
Bin 2: 1943-1948	9	176.9	121.0	501
Bin 3: 1949-1954	18	232.0	272.0	496
Bin 4: 1955-1960	23	365.0	425.0	495
Bin 5: 1961-1966	11	473.2	486.0	600
NS3				
Bin 1: 1928-1944	11	52.6	50.0	139
Bin 2: 1945-1950	8	88.4	96.0	214
Bin 3: 1951-1956	24	116.1	99.5	292
Bin 4: 1957-1967*	10	70.2	17.0	328

*For NS3, only four bins were created

Table 4.18. Annual measurements stratified by birth year: electrician SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: Prior to 1955	91	88.3	3.0	1691
Bin 2: 1955 or later	95	129.6	4.0	1628
NS2				
Bin 1: Prior to 1955	83	61.1	18.0	447
Bin 2: 1955 or later	49	114.6	84.0	439
NS3				
Bin 1: Prior to 1950	56	32.2	1.0	232
Bin 2: 1950 or later	48	60.1	8.0	439

Table 4.19. Daily measurements stratified by birth year: electrician SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: Prior to 1955	262	30.7	2.0	872
Bin 2: 1955 or later	282	43.7	2.0	1266
NS2				
Bin 1: Prior to 1955	258	19.7	5.5	289
Bin 2: 1955 or later	160	35.1	17.0	317
NS3				
Bin 1: Prior to 1950	172	10.6	1.0	164
Bin 2: 1950 or later	130	22.2	2.0	271

Table 4.20. Annual measurements stratified into six bins by birth year: electrician SEG

Bin	No. of Workers	Mean Exposure (mrem)	Median Exposure (mrem)	Peak Exposure (mrem)
NS1				
Bin 1: 1930-1940	8	69.8	48.0	337
Bin 2: 1941-1946	18	40.9	2.0	202
Bin 3: 1947-1952	47	112.7	5.0	1691
Bin 4: 1953-1958	57	140.0	4.0	1628
Bin 5: 1959-1964	42	69.4	4.5	379
Bin 6: 1965-1968	14	203.6	2.5	1511
NS2				
Bin 1: 1925-1935	5	45.0	5.0	133
Bin 2: 1936-1941	13	35.9	11.0	153
Bin 3: 1942-1947	22	80.7	35.0	260
Bin 4: 1948-1953	36	67.1	20.0	447
Bin 5: 1954-1959	35	108.1	84.0	439
Bin 6: 1960-1967	21	96.0	51.0	416
NS3				
Bin 1: 1931-1939	16	16.7	1.0	148
Bin 2: 1940-1945	15	18.7	12.0	75
Bin 3: 1946-1951	38	58.7	24.5	350
Bin 4: 1952-1957	16	73.9	21.0	439
Bin 5: 1958-1967*	19	38.4	0.0	297

*For NS3, only five bins were created

4.3. Results

4.3.1. Grouping Measurements by Time Intervals

Missing daily radiation exposure data within the pipefitting SEG were imputed using three separate models. The models differed only by the variable used to describe sample collection year.

Estimated SEG Mean

The estimated MI mean was similarly accurate in all three models within NS1; at very high percentages of missing data (99.0-99.95%), the model using 10-year time intervals produced estimates of the mean that had the least bias (Table 4.21). The models using a time interval also produced estimates of the mean that were closer to the true mean within NS2 and NS3, particularly when the percentage of missing data was very high. Although the width of the 95% CI varied, the models using time intervals were observed to have narrower confidence intervals.

Estimated SEG Median

NS1: The estimated MI medians within NS1 were very close to the true median in all trials; however, the model using 10-year time intervals produced estimates of the median that had the most bias (Table 4.21). Within NS2, no difference in the estimated MI medians was observed until the percentage of missing data reached 99.90%. At the highest percentages of missing data, the model using no intervals produced estimates of the median that had the least bias. Within NS3, no difference in the estimated MI medians was observed until the percentage of missing data reach 99.95%.

Estimated Imputation Variance

Using a model with a 5-year or 10-year time interval generally underestimated the true variance of the dataset within NS1, particularly at higher percentages of missing data (Table 4.21). Within NS2, using a model with a 5-year or 10-year time interval was shown to fluctuate between over- and underestimating the true variance, with the most severe underestimations observed at the highest percentages of missing data. For NS3, using models with a time interval produced estimates of the variance in all trials that were underestimates of the true variance.

Table 4.21. Performance of MI by percent missing data for pipefitting SEG: NS1 (mean = 12.5 mrem, median = 1.0 mrem, SE = 55.6)

% Missing	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
All available years						
50.0	1.2	(0, 1267)	0.0	3.32E+03	0.3	57.6
90.0	1.1	(0, 1356)	0.0	3.87E+03	0.7	62.2
95.0	-0.1	(0, 131)	-0.2	3.66E+03	3.0	60.6
99.0	3.0	(0, 1467)	0.0	4.49E+03	1.9	67.0
99.90	-5.4	(0, 53)	-0.4	5.30E+02	9.9	23.3
99.95	9.1	(0, 160)	0.0	4.91E+03	46.9	70.5
5-year intervals						
50.0	0.2	(0, 122)	-1.0	3.09E+03	0.1	55.6
90.0	-2.3	(0, 101)	0.0	2.14E+03	0.8	46.3
95.0	-1.2	(0, 120)	-0.4	3.06E+03	2.7	55.4
99.0	3.6	(0, 168)	0.0	6.03E+03	1.8	77.7
99.90	-8.6	(0, 34)	-0.6	2.34E+02	0.3	15.3
99.95	-10.9	(0, 10)	-0.8	1.90E+01	0.0	4.4
10-year intervals						
50.0	-0.9	(0, 116)	-0.8	2.84E+03	0.1	53.3
90.0	1	(0, 124)	-1.0	3.17E+03	2.8	56.3
95.0	-2.5	(0, 113)	-0.6	2.76E+03	4.3	52.6
99.0	-2.1	(0, 95)	-0.4	1.85E+03	14.1	43.2
99.90	-5.2	(0, 75)	-0.6	1.20E+03	3.5	34.7
99.95	-6.3	(0, 66)	-1.0	9.31E+02	9.7	30.7

Table 4.22. Performance of MI by percent missing data for pipefitting SEG: NS2 (mean = 18.4 mrem, median = 1.0 mrem, SE = 85.6)

% Missing	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
All available years						
50.0	1.5	(0, 189)	0.0	7.44E+03	0.9	86.3
90.0	6.9	(0, 249)	0.0	1.30E+04	21.1	114.3
95.0	6.2	(0, 250)	0.0	1.32E+04	43.0	115.0
99.0	8.4	(0, 271)	0.0	1.55E+04	14.3	124.6
99.90	13.0	(0, 235)	0.0	1.08E+04	53.5	104.1
99.95	25.0	(0, 244)	-0.6	2.28E+04	563.1	153.2
5-year intervals						
50.0	-1.0	(0, 168)	0.0	5.90E+03	1.3	76.8
90.0	-2.4	(0, 164)	0.0	5.66E+03	4.2	75.3
95.0	3.4	(0, 212)	0.0	9.36E+03	39.5	97.0
99.0	0.2	(0, 204)	0.0	8.91E+03	13.1	94.5
99.90	-5.2	(0, 96)	-1.0	1.74E+03	24.9	42.1
99.95	-7.5	(0, 91)	-1.0	1.64E+03	35.0	41.0
10-year intervals						
50.0	0.6	(0, 190)	0.0	7.58E+03	1.7	87.1
90.0	-3.5	(0, 156)	0.0	5.15E+03	17.9	71.9
95.0	-0.4	(0, 198)	0.0	8.34E+03	52.8	91.7
99.0	0.6	(0, 220)	0.0	1.05E+04	13.2	102.5
99.90	-7.4	(0, 72)	-0.6	9.21E+02	40.4	31.1
99.95	-4.3	(0, 117)	-1.0	2.67E+03	61.7	52.3

Table 4.23. Performance of MI by percent missing data for pipefitting SEG: NS3 (mean = 13.2 mrem, median = 0.0 mrem, SE = 95.0)

% Missing	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
All available years						
50.0	1.5	(0, 202)	0.0	9.10E+03	0.4	95.4
90.0	1.8	(0, 173)	0.0	6.49E+03	4.5	80.6
95.0	1.1	(0, 168)	0.0	6.14E+03	0.3	78.4
99.0	1.9	(0, 160)	0.0	5.36E+03	100.8	74.1
99.90	5.4	(0, 143)	0.0	3.99E+03	29.1	63.4
99.95	39.8	(0, 320)	0.6	1.83E+04	194.9	136.2
5-year intervals						
50.0	0.5	(0, 198)	0.0	8.83E+03	2.9	94.0
90.0	-2.0	(0, 138)	0.0	4.17E+03	1.1	64.6
95.0	1.5	(0, 175)	0.0	6.63E+03	19.6	81.6
99.0	-3.5	(0, 113)	0.0	2.78E+03	2.1	52.8
99.90	3.1	(0, 126)	0.0	3.11E+03	4.8	55.8
99.95	-5.7	(0, 73)	0.0	1.11E+03	1.3	33.4
10-year intervals						
50.0	-0.9	(0, 184)	0.0	7.68E+03	2.0	87.6
90.0	-2.0	(0, 139)	0.0	4.28E+03	1.2	65.4
95.0	-2.7	(0, 130)	0.0	3.69E+03	6.9	60.8
99.0	-3.1	(0, 118)	0.0	3.05E+03	1.2	55.3
99.90	-2.4	(0, 100)	0.0	2.05E+03	8.5	45.4
99.95	-4.6	(0, 78)	0.0	1.25E+03	0.5	35.3

4.3.2. Variation in Number of Samples Collected

In this section, the total number of samples collected within an SEG is varied by the number of workers sampled and the number of samples collected per worker. Two separate exercises were examined: one in which only those measurements selected to be sampled were used to calculate exposure ("subset only") and another in which those measurements not sampled were imputed and then included in the analysis ("imputed").

Pipefitting

Estimated SEG Mean

As shown in Tables 4.24-4.26, the subset only and the imputed scenarios generally performed similarly to one another within a trial in estimating the mean. The notable exception occurred when the percentage of workers sampled was lowered to 5.0%. In those specific trials, the subset only scenarios performed better than the imputed scenarios. The width of the 95% CI varied between the two scenarios, with no discernable pattern. Increasing the number of samples collected per worker within a percentage of workers sampled did not always improve the accuracy of the estimated mean and in some cases greatly worsen the accuracy. Varying the number of samples and percentage of workers sampled had varying effects across the yards. Increasing the number of samples collected per worker also appeared to increase the width of the 95% CI.

Estimated SEG Median

As was also observed with the estimated mean, the subset only and the imputed scenarios performed similarly to one another within a trial in estimating the median (Tables 4.24-4.26). Increasing the number of samples collected per worker within a percentage of workers sampled appeared to vary in its effect on the accuracy of the estimated median. A strong pattern was not observed.

Estimated Imputation Variance

The total estimated variance within a trial was higher in the imputed scenario in approximately one-half or more of the trials within each yard (Tables 4.24-4.26). In some

trials, increasing the number of samples collected per worker increased the estimated variance, but this was not always observed. However, collecting only one sample per worker resulted in underestimations of the true variance at almost all percentages of workers sampled. When the percentage of workers sampled was set at 20% or less, the between-imputation variance for the MI trials increased significantly.

Welding

Estimated SEG Mean

In general, the accuracy of the estimated means was worse for the welding SEG compared to the pipefitting SEG. Greater differences in bias were also observed within a trial when looking at the welding SEG, particularly at lower percentages of workers sampled (Tables 4.27-4.29). The 95% CI widths were also more variable.

Estimated SEG Median

Again, the overall accuracy of the estimated medians was worse for the welding SEG compared to the pipefitting SE (Tables 4.27-4.29). Increasing the number of samples collected per worker within a percentage of workers sampled appeared to vary in its effect on the accuracy of the estimated median. A strong pattern was not observed.

Estimated Imputation Variance

The majority of the estimated variances were underestimates of the true variance (Tables 4.27-4.29). Increasing the number of samples collected per worker within a percentage of workers sampled generally increased the estimated variance, particularly within the imputed

scenarios. However, collecting the same number of total samples, but from a smaller percentage of the population, decreased the estimated variance. Unlike with the pipefitting SEG, the between-imputation variance did not appear to follow a pattern as the percentage of workers sampled varied.

Table 4.24. Performance of MI by sampling plan for pipefitting SEG: NS1 (mean = 43.1, median = 3.0, SE = 104.9)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-4.0	(0, 216)	-2.0	--	--	90.1
	Imputed	-6.0	(0, 248)	-1.6	1.15E+04	58.2	107.5
100% (2)	Subset only	-5.9	(0, 252)	-1.0	--	--	109.3
	Imputed	-7.6	(0, 185)	-1.2	5.82E+03	2.8	76.3
50% (1)	Subset only	-12.3	(0, 217)	-2.0	--	--	94.8
	Imputed	-10.9	(0, 212)	-1.0	8.32E+03	74.4	91.7
50% (2)	Subset only	4.8	(0, 244)	2.0	--	--	99.8
	Imputed	-4.6	(0, 234)	-0.2	9.95E+03	43.1	100.0
50% (4)	Subset only	11.1	(0, 293)	5.0	--	--	121.6
	Imputed	17.9	(0, 373)	5.0	2.53E+04	41.2	159.1
20%(1)	Subset only	-24.8	(0, 141)	-3.0	--	--	62.3
	Imputed	-19.6	(0, 187)	-3.0	6.85E+03	59.6	83.2
20% (2)	Subset only	9.6	(0, 233)	3.5	--	--	92.1
	Imputed	12.8	(0, 368)	4.8	2.52E+04	166.9	159.2
20% (4)	Subset only	7.2	(0, 290)	1.0	--	--	122.1
	Imputed	27.1	(0, 452)	3.2	3.72E+04	538.1	194.6
10% (1)	Subset only	-2.9	(0, 161)	-2.0	--	--	61.5
	Imputed	-5.5	(0, 193)	-2.0	6.31E+03	15.2	79.5
10% (2)	Subset only	1.8	(0, 192)	11.5	--	--	75.2
	Imputed	-2.3	(0, 171)	10.2	4.41E+03	26	66.6
10% (4)	Subset only	23.6	(0, 268)	4.0	--	--	102.9
	Imputed	18.9	(0, 325)	5.8	1.78E+04	133.7	134
5% (1)	Subset only	2.2	(0, 123)	7.0	--	--	39.4
	Imputed	-9.6	(0, 128)	3.0	2.26E+03	57.3	48.2
5% (2)	Subset only	5.3	(0, 235)	2.0	--	--	95.3
	Imputed	17.6	(0, 233)	13.0	8.38E+03	189.7	92.8
5% (4)	Subset only	9.1	(0, 456)	8.0	--	--	206.0
	Imputed	28.6	(0, 430)	0.8	3.25E+04	681.1	182.6

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

Table 4.25. Performance of MI by sampling plan for pipefitting SEG: NS2 (mean = 52.2, median = 22.0, SE = 97.3)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-10.1	(0, 221)	-9.0	--	--	90.8
	Imputed	-13.9	(0, 169)	-9.6	4.38E+03	7	66.3
100% (2)	Subset only	-6.8	(0, 213)	-5.0	--	--	84.9
	Imputed	-7.5	(0, 197)	-3.6	6.02E+03	4	77.6
50% (1)	Subset only	-5.2	(0, 185)	-9.0	--	--	70.1
	Imputed	-2.9	(0, 226)	-7.4	7.99E+03	80	89.9
50% (2)	Subset only	-7.9	(0, 201)	-2.0	--	--	79.5
	Imputed	-0.4	(0, 218)	1.6	7.10E+03	23.8	84.4
50% (4)	Subset only	4.9	(0, 239)	5.0	--	--	92.4
	Imputed	3.9	(0, 256)	4.4	1.03E+04	8.2	101.7
20%(1)	Subset only	-0.6	(0, 173)	4.0	--	--	61.5
	Imputed	-20.1	(0, 127)	-0.8	2.33E+03	7.2	48.4
20% (2)	Subset only	-0.5	(0, 333)	-4.5	--	--	143.5
	Imputed	-6.6	(0, 200)	-3.2	6.13E+03	33.8	78.6
20% (4)	Subset only	8.2	(0, 285)	12.5	--	--	114.3
	Imputed	-0.8	(0, 223)	5.0	7.57E+03	44.5	87.3
10% (1)	Subset only	6.2	(0, 198)	4.0	--	--	70.8
	Imputed	3.2	(0, 253)	-1.8	1.01E+04	22	100.7
10% (2)	Subset only	-1.9	(0, 171)	10.0	--	--	61.5
	Imputed	-2.6	(0, 179)	5.0	4.23E+03	70.3	65.7
10% (4)	Subset only	-4.9	(0, 284)	4.5	--	--	120.7
	Imputed	-10.8	(0, 162)	4.6	3.76E+03	14.4	61.4
5% (1)	Subset only	-4.7	(0, 131)	0.0	--	--	42.2
	Imputed	-2.5	(0, 147)	10.6	2.36E+03	55.4	49.3
5% (2)	Subset only	-16.3	(0, 186)	-4.5	--	--	76.5
	Imputed	-12.2	(0, 145)	-2.8	2.63E+03	170.5	53.3
5% (4)	Subset only	-13.1	(0, 322)	-13.0	--	--	144.1
	Imputed	1.1	(0, 368)	-12.6	2.54E+04	212.7	160.2

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

Table 4.26. Performance of MI by sampling plan for pipefitting SEG: NS3 (mean = 31.0, median = 3.0, SE = 66.7)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-2.3	(0, 190)	-1.0	--	--	82.7
	Imputed	2.2	(0, 177)	0.0	5.36E+03	20.6	73.4
100% (2)	Subset only	-1.9	(0, 150)	-1.0	--	--	61.7
	Imputed	-1.5	(0, 157)	-0.8	5.36E+03	20.6	73.4
50% (1)	Subset only	0.8	(0, 158)	0.5	--	--	64.7
	Imputed	5.5	(0, 152)	0.4	3.49E+03	5.5	59.1
50% (2)	Subset only	7.4	(0, 138)	0.0	--	--	51.3
	Imputed	7.6	(0, 187)	0.2	5.79E+03	8.1	76.1
50% (4)	Subset only	4.0	(0, 167)	2.0	--	--	67.5
	Imputed	1.5	(0, 165)	0.6	4.61E+03	3.3	67.9
20%(1)	Subset only	3.6	(0, 162)	-0.5	--	--	65.2
	Imputed	8.3	(0, 165)	6.0	3.59E+03	486.6	64.6
20% (2)	Subset only	-2.4	(0, 155)	2.5	--	--	65
	Imputed	7.8	(0, 189)	5.2	5.54E+03	305.7	76.9
20% (4)	Subset only	9.2	(0, 168)	4.0	--	--	65.2
	Imputed	13.6	(0, 186)	6.0	5.22E+03	14.1	72.4
10% (1)	Subset only	-5.7	(0, 128)	-1.0	--	--	52.7
	Imputed	8.3	(0, 226)	6.0	8.52E+03	486.6	95.4
10% (2)	Subset only	3.6	(0, 195)	5.0	--	--	81.9
	Imputed	5.8	(0, 177)	4.8	5.11E+03	13.3	71.6
10% (4)	Subset only	4.4	(0, 173)	1.5	--	--	70.5
	Imputed	3.1	(0, 174)	1.6	5.04E+03	81.6	71.7
5% (1)	Subset only	-8.6	(0, 60)	8.0	--	--	19.7
	Imputed	-16.7	(0, 51)	4.0	3.52E+02	0.4	18.8
5% (2)	Subset only	-18.2	(0, 91)	0.5	--	--	40.1
	Imputed	-21.2	(0, 44)	0.6	3.02E+02	8.6	17.7
5% (4)	Subset only	8.9	(0, 196)	3.5	--	--	79.9
	Imputed	10.0	(0, 152)	17.2	2.96E+03	122.5	55.7

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

Table 4.27. Performance of MI by sampling plan for welding SEG: NS1 (mean = 76.1, median = 17.0, SE = 157.2)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-39.8	(0, 297)	-13.0	--	--	133.0
	Imputed	-30.1	(0, 268)	-13.4	1.28E+04	62.3	113.5
100% (2)	Subset only	-22.2	(0, 333)	-9.0	--	--	142.4
	Imputed	-23.5	(0, 312)	-11.2	1.75E+04	36.2	132.3
50% (1)	Subset only	-31.7	(0, 349)	-17.0	--	--	155.4
	Imputed	-30.3	(0, 289)	-17.0	1.50E+04	325.5	124.1
50% (2)	Subset only	-19.5	(0, 262)	-3.5	--	--	104.9
	Imputed	-19.1	(0, 238)	4.8	8.49E+03	96.8	92.7
50% (4)	Subset only	14.4	(0, 403)	17.0	--	--	159.8
	Imputed	22.3	(0, 437)	21.8	2.99E+04	21.3	173.1
20%(1)	Subset only	6.1	(0, 313)	-17.0	--	--	117.9
	Imputed	8.1	(0, 444)	-15.8	3.15E+04	1921.8	183.8
20% (2)	Subset only	-40.0	(0, 299)	-9.0	--	--	134.2
	Imputed	-41.5	(0, 154)	-8.8	3.65E+03	88.8	61.3
20% (4)	Subset only	-1.8	(0, 390)	8.0	--	--	161.2
	Imputed	23.8	(0, 562)	5.2	5.40E+04	1406.7	236.0
10% (1)	Subset only	-14.9	(0, 144)	-17.0	--	--	42.4
	Imputed	-5.1	(0, 306)	-11.8	1.41E+04	340.5	120.3
10% (2)	Subset only	-19.7	(0, 213)	8.0	--	--	80.1
	Imputed	-21.4	(0, 211)	5.0	6.15E+03	184.2	79.8
10% (4)	Subset only	50.4	(0, 403)	13.0	--	--	141.4
	Imputed	9.7	(0, 482)	4.2	4.03E+04	442.6	202.1
5% (1)	Subset only	-58.7	(0, 63)	-17.0	--	--	23.3
	Imputed	-58.8	(0, 59)	-17.0	4.61E+02	1.8	21.5
5% (2)	Subset only	95.9	(0, 243)	-4.0	--	--	36.3
	Imputed	64.1	(0, 622)	-7.0	6.00E+04	511.4	246.1
5% (4)	Subset only	-16.1	(0, 140)	1.5	--	--	41.3
	Imputed	-27.9	(0, 185)	-3.8	4.59E+03	263.2	70.0

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

Table 4.28. Performance of MI by sampling plan for welding SEG: NS2 (mean = 89.4, median = 54.0, SE = 94.7)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-12.7	(0, 259)	-4.0	--	--	93.2
	Imputed	-4.8	(0, 258)	3.9	7.87E+03	29.9	88.9
100% (2)	Subset only	-3.7	(0, 250)	-2.0	--	--	84.3
	Imputed	-1.8	(0, 262)	13.0	7.86E+03	91.7	89.3
50% (1)	Subset only	-28	(0, 212)	-28.0	--	--	77.2
	Imputed	-1.1	(0, 270)	19.4	8.42E+03	219.8	93.2
50% (2)	Subset only	-0.1	(0, 264)	10.0	--	--	89.3
	Imputed	4.5	(0, 260)	22.4	7.22E+03	9.1	85.0
50% (4)	Subset only	8.0	(0, 279)	23.0	--	--	92.7
	Imputed	8.1	(0, 275)	18.8	8.24E+03	16.3	90.9
20%(1)	Subset only	-40.7	(0, 156)	-21.0	--	--	55.0
	Imputed	-49.5	(0, 134)	-50.4	2.30E+03	19.8	48.2
20% (2)	Subset only	-13.5	(0, 256)	-29.0	--	--	92.2
	Imputed	-15.6	(0, 247)	-28.6	7.79E+03	15.7	88.4
20% (4)	Subset only	4.1	(0, 244)	28.0	--	--	77.2
	Imputed	11.1	(0, 273)	29.2	7.41E+03	280.9	88.0
10% (1)	Subset only	-45.0	(0, 219)	-45.0	--	--	89.3
	Imputed	-50.9	(0, 149)	-47.4	3.10E+03	71.8	56.4
10% (2)	Subset only	39.7	(0, 310)	74.5	--	--	92.7
	Imputed	27.2	(0, 295)	55.3	8.23E+03	84.4	91.3
10% (4)	Subset only	-9.0	(0, 188)	-15.5	--	--	55.0
	Imputed	-12.1	(0, 258)	-16.0	8.50E+03	23.3	92.3

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

Table 4.29. Performance of MI by sampling plan for welding SEG: NS3 (mean = 28.8, median = 9.5, SE = 39.7)

% Workers Sampled (No. samples per worker)	Scenario*	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
100% (1)	Subset only	-7.7	(0, 100)	-8.0	--	--	40.2
	Imputed	-3.6	(0, 108)	-4.9	1.78E+03	4.8	42.3
100% (2)	Subset only	-5.9	(0, 88)	-3.0	--	--	33.4
	Imputed	-6.0	(0, 86)	-2.1	1.04E+03	3.8	32.4
50% (1)	Subset only	-16.2	(0, 49)	-4.5	--	--	19.0
	Imputed	-16.8	(0, 48)	-6	3.47E+02	5.6	18.8
50% (2)	Subset only	7.1	(0, 128)	8.5	--	--	47.2
	Imputed	5.2	(0, 123)	4.8	2.05E+03	21.8	45.5
50% (4)	Subset only	6.2	(0, 120)	9.0	--	--	43.7
	Imputed	10.1	(0, 129)	15.6	2.12E+03	0.9	46
20%(1)	Subset only	-13.0	(0, 53)	-0.5	--	--	19.0
	Imputed	-18.8	(0, 37)	-7.7	1.91E+02	1.3	13.9
20% (2)	Subset only	5.2	(0, 116)	6.0	--	--	42.1
	Imputed	11.4	(0, 120)	35.5	1.64E+03	14.4	40.8
20% (4)	Subset only	1.8	(0, 94)	12.0	--	--	32.7
	Imputed	3.4	(0, 97)	12.7	1.08E+03	14.1	33.2
10% (1)	Subset only	21.7	(0, 159)	14.5	--	--	55.8
	Imputed	22.1	(0, 151)	14.7	2.60E+03	18.2	51.2
10% (2)	Subset only	-6.0	(0, 79)	-4.0	--	--	29.0
	Imputed	-10.2	(0, 70)	-3.0	6.41E+02	60.3	26.7
10% (4)	Subset only	-1.4	(0, 101)	-1.0	--	--	37.8
	Imputed	0.1	(0, 102)	1.9	1.41E+03	15.1	37.8

*Subset only = non-sampled measurements were not analyzed; Imputed = non-sampled measurements were imputed

4.3.3. Exploring Additional Exposure Covariates

In the third and final section, additional information on the work population, beyond job title, will be considered in defining SEGs. Specifically, the variables *birth year* and *sample collection quarter* were examined within the pipefitting SEG.

Birth year

Estimated SEG Mean

As shown in Table 4.30, removing the variable *birth year* resulted in estimations of the mean that were approximately 11-21 mrem away from the true mean. Removing the variable *education level* resulted in estimations of the mean that were approximately 5-14 mrem away from the true mean. When the variable *race* was removed from the models used for NS1 and NS3, the estimations of the mean were approximately 7-12 mrem away from the true mean (race variable was not added to model in NS2). Removing any one of the variables from the model resulted in a narrower confidence interval compared to the model that contained all the variables when examining NS1 or NS3; when using the data for NS2, removing any one of the variables resulted in a wider confidence interval.

Estimated SEG Median

Removing the variable *birth year* resulted in estimations of the median that were approximately 1-7 mrem away from the true median (Table 4.30). When the variable *education level* was removed, estimations of the median were approximately 1-3 mrem away from the true median. When the variable *race* was removed from the models used for NS1

and NS3, the estimations of the median were approximately 5 mrem away from the true median.

Estimated Imputation Variability

All but one trial resulted in variance estimations that were underestimates of the true variance (Table 4.30). Removing the variable *birth year* resulted in the worst estimations of the true variance. For NS1 and NS3, removing the variable *birth year* resulted in the lowest variance estimates; for NS2, removing *birth year* actually resulted in an overestimation of the true variance. For NS1 and NS3, the between-imputation variance was also the lowest when *birth year* was removed.

Table 4.30. Performance of MI by variable removed from model: pipefitting SEG

Variable Removed from Model	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (mean = 142.6 mrem, median = 27.0 mrem, SE = 260.7)						
None	-5.6	(0, 637)	-0.6	6.49E+04	266.7	255.3
Birth year	-20.7	(0, 553)	-6.6	4.82E+04	229.4	220.3
Education	-5.1	(0, 634)	0.7	6.34E+04	621.2	253.3
Race	-12.1	(0, 580)	4.7	5.23E+04	278.0	229.3
NS2 (mean = 192.1, median = 122.5 mrem, SE = 246.0)						
None	-23.3	(0, 599)	-13.2	4.82E+04	37.5	219.6
Birth year	18.7	(0, 801)	2.1	8.98E+04	715.0	301.2
Education	-8.3	(0, 655)	-2.7	5.73E+04	522.2	240.8
NS3 (122.7 mrem, median = 31.0 mrem, SE = 155.8)						
None	-4.2	(0, 414)	1.6	2.27E+04	74.5	151.0
Birth year	-11.2	(0, 394)	0.8	2.08E+04	26.5	144.4
Education	-13.9	(0, 393)	-3.4	2.09E+04	171.4	145.4
Race	-6.7	(0, 407)	5.6	2.21E+04	63.9	148.8

Sample Collection Quarter

Estimated SEG Mean

The models in which only the sample collection year variable was used resulted in estimations of the mean that were closest to the true mean at two of the yards (NS1 and NS2) (Table 4.31). At all three yards, the models in which only the sample collection quarter variable was used resulted in the most biased estimates of the mean. The width of 95% confidence intervals varied by yard.

Estimated SEG Median

There was no observed difference between the estimated median and the true median within all trials for all three yards (Table 4.31).

Estimated Imputation Variability

The estimated total variances associated with all three models were overestimates of the true variance for almost all of the trials (Table 4.31). The model that included sample collection quarter only had the highest overestimate of the total variance at two of the three yards (NS1 and NS2).

Table 4.31. Performance of MI by sample collection date variable(s) included: pipefitting SEG

Sample Collection Variable(s) Included	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 (mean = 12.5 mrem, median = 1.0 mrem, SE = 55.6)						
Both	0.9	(0, 127)	0.0	3.36E+03	0.1	58.0
Year Only	0.7	(0, 124)	0.0	3.21E+03	0.0	56.7
Quarter Only	2.6	(0, 133)	0.0	3.65E+03	0.6	60.4
NS2 (mean = 18.4 mrem, median = 1.0 mrem, SE = 85.6)						
Both	3.5	(0, 210)	0.0	9.24E+03	0.4	96.1
Year Only	1.2	(0, 189)	0.0	7.50E+03	1.1	86.6
Quarter Only	6.7	(0, 227)	0.0	1.06E+04	3.7	103.1
NS3 (mean = 13.2 mrem, median = 0.0 mrem, SE = 95.0)						
Both	1.7	(0, 202)	0.0	9.11E+03	2.2	95.5
Year Only	3.6	(0, 220)	0.0	1.07E+04	1.9	103.7
Quarter Only	4.5	(0, 216)	0.0	1.03E+04	0.8	101.4

4.4. Discussion

4.4.1. Grouping Measurements by Time Intervals

In this section, the number of radiation measurements available over time within an SEG was varied and the missing data were imputed using one of three ways to describe sample collection date: using each available year, combining years into 5-year time intervals, and combining years into 10-year time intervals. The objectives of this section were to examine the ability of using a limited number of measurements per year to characterize the exposure profile of an SEG through a multiple imputation approach and to explore the potential of grouping measurements into larger time intervals to assist in estimating exposures within an SEG.

When estimating exposure levels within an SEG, the number of available measurements will likely be significantly fewer than were available for the entire population. This is particularly

true if the number of SEGs developed by the hygienist is large. Investigators attempting to estimate exposure levels through modeling may therefore be concerned about sample sizes, especially if there is a large amount of missing exposure data. Combining data into broader intervals may be a helpful solution when one of the model covariates has many levels and only a few samples per level (sample collection year is one good example of this). By grouping data points together into intervals, each covariate level now has a larger sample size. This may improve the performance of the desired modeling technique.

At lower percentage of missing data, all three models performed similarly well. As the percentage of missing data increased, the performance of the models did drop. However, the model using the 10-year time interval was shown to perform the best in estimating the SEG mean exposure level. Combining data into broader intervals is therefore one potential solution. Investigators should be aware, however, that grouping data into bins will likely result in an underestimation of the total variance. Whether this is a large concern will depend on the ultimate goal of the analysis.

4.4.2. Variation in Number of Samples Collected

In this section, the total number of samples collected within an SEG is varied by the number of workers sampled and the number of samples collected per worker. The first objective was to explore how various sampling strategies affect the ability to accurately estimate the true SEG exposure profile. The second objective was to investigate whether an MI approach can assist in developing more accurate exposure estimates.

In total, 14 separate sampling plans were explored. In general, using only the subset of sampled workers to characterize the exposure level of the entire SEG worked fairly well, particularly when estimating exposures for the pipefitting SEG. Surprisingly, strong patterns were not observed when the percentage of workers sampled or the number of samples collected per worker was varied. Imputing the missing, or "not sampled," measurements may improve the overall exposure estimates slightly but was surprisingly not as beneficial as might be expected. While the performance of the multiple imputation approach varied by trial, the overall results generally agreed with those obtained using the subset of sampled measurements only.

A few comments can be made from these analyses. First, collecting multiple samples per worker, when possible, is recommended. While the estimations of exposure mean and median did not always improve, collecting multiple measurements on each sampled worker did generally result in better estimations of the true variance. This is in agreement with earlier work by Rappaport in which he recommended a sampling plan in which every worker is monitored at least twice in order to understand the within-worker variability (Rappaport, 1991). Collecting only one sample per worker may in fact heavily underestimate the variance. Second, collecting samples from a larger percentage of workers did not necessarily result in improved estimates of exposure. For example, collecting multiple samples on 20% of the work population generally did not result in exposure estimates that were much more biased than collecting one sample on 50% of the population. This can be a reassuring result, as it is often very difficult, if not impossible, for a hygienist to collect samples on the entire work population within an SEG. Third, both scenarios generally performed better when using

the data from the pipefitting SEG, as compared to the welding SEG. This suggests that the homogeneity of exposures within an SEG likely varies by exposure group and by the work tasks performed by that exposure group.

This is, of course, part of a larger concern surrounding SEGs. Collecting data on 20% of the SEG population may be appropriate, but only if that 20% of workers is truly similar to the unmeasured 80% of workers. Although the subset and multiple imputation scenarios performed equally in many cases, they often both produced mean exposure estimates that were more than 10.0 mrem away from the true mean. It is possible that such bias is a result of a heterogeneous exposure group.

4.4.3. Exploring Additional Exposure Covariates

Birth Year

The relationship between birth year and mean exposure level was observed for all three SEGs and all three shipyards. This suggests that birth year likely has some indirect effect on worker exposure level. As discussed above, there is a possibility that birth year is a surrogate for work task, or at the least, a more specific description of the worker's job title. Skilled trades are known to have various stages in their respective fields. Generally, these stages are referred to as apprentice, journeyman, and master. At each stage, the work responsibilities change. A master in a skilled trade is often responsible for training the apprentice and overseeing the journeyman's work. Thus, it was not surprising to see that the older workers, who were more likely to be masters, had lower mean and peak exposures compared to the

younger workers. Exposure information such as this may often be hiding within other variables; if these variables are not difficult to collect, they can be valuable sources of information.

Removing any one of the three tested variables (birth year, education level, and race) resulted in poorer performing models, and removing birth year sometimes resulted in the worst performing ones. The fact that removing birth year resulted in the lowest estimates of variance suggests that the within-SEG variability is at least partially due to the birth year variable. In fact, when examined closer, the education level and race within an SEG were much more homogenous than compared to birth year. Thus, birth year, for this work population, is an important variable in include in exposure models.

Given the heterogeneous exposures by birth year, the appropriateness of using job title to define the SEGs has to be reconsidered. Combining all pipefitters, for example, into one SEG may not be correct; the exposure levels within the pipefitting trade may be too varied by job task to be combined. This is a particularly important consideration for occupational exposure studies in which SEGs are often defined based on broader categories. Prior to defining exposure groups, as much should be learned about the industry, job title, and work tasks as possible.

Sample Collection Quarter

Surprisingly, sample collection quarter did not affect the performance of the model very much. It is clear that sample collection year is a very important variable to include, but the quarter did not appear to add much additional information. While it is still possible that the variable *quarter* contains some information about the ship overhaul and maintenance schedule, this is impossible to confirm without additional information.

4.5. Conclusion

The analyses in Chapter 4 allowed for a better understanding of the factors that influence the homogeneity of an SEG and thus influence the ability to accurately characterize the exposures of the work force. Industrial hygienists and researchers alike are commonly faced with many decisions when building a sampling plan. The goal becomes to design a plan that captures enough variability to answer the intended study questions. Understanding the determinants of exposure for a particular workforce can assist in assigning workers to more homogeneous SEGs. In addition, when deciding which approach to take in addressing missing data, the ultimate goals of the study should be considered; the relative importance of factors such as unbiased exposure estimates versus more a more accurate estimate of variance should be weighed.

CHAPTER 5: DEVELOPING EXPOSURE ESTIMATES USING SURROGATE DATA

5.1. Introduction

5.1.1. Combining Exposure Data from Multiple Facilities

The type and quality of exposure data often vary between epidemiologic studies; this can also be true of different facilities within the same industry and even within one location over prolonged periods of time (Checkoway et al. 2004). If measurement data for the location under study are sparse or strongly biased in some way, the researcher may consider combining the existing data with available measurements from similar but separate facilities. In this sense, the exposure data from other facilities can be considered surrogate data for the location of interest. The benefit of this strategy is an increase in the number of available exposure measurements from which to develop an exposure profile for the study population of interest. The assumption being made when employing such a technique, however, is that the exposure levels are similar between sites, particularly within job titles or SEGs. If this assumption is incorrect, there is the potential for exposure misclassification. This concern increases if the type of exposure data (quantitative exposure measurements, qualitative exposure rankings, etc.) varies between facilities.

In this study, exposure data that are similar in quality are available for three separate shipyards. The job titles were observed to be largely the same between yards and the work tasks and materials used are likely to be similar, since all procedures conformed to the same military specifications. Thus, it may be initially assumed that exposure levels between yards are comparable. If exposure data were limited at one of the shipyards, this assumption would

then allow for a combining of available exposure data from each yard to better characterize the exposure profile of the population at the shipyard of interest. However, exposure variability within and between shipyards may result in differences between the yards that can bias the estimates. Thus, prior to combining exposure data from multiple locations, an attempt should be made to understand the observed exposure trends within and between the yards. Differences between the yards may also vary based on the calendar year and time span of the study period of interest, as well as on the size of the study population.

This chapter will compare exposure levels between the three shipyards at both the population and SEG level and during different time periods. Exposure data from one yard at a time will then be assigned to be missing; the remaining exposure data available from each of the yards will be combined and used to impute the missing data.

5.1.2. Specific Aims

The specific aims of this chapter are to compare between shipyards the exposure profile of naval shipyard workers during various time periods and to test the performance of a multiple imputation approach in estimating exposure levels when surrogate exposure data are used.

5.1.3. Focused Research Objectives

In order to test the performance of the multiple imputation method, artificially missing exposure data were generated from each complete dataset. In this chapter, data were randomly assigned to be missing at varying proportions. The analyses performed in this chapter are summarized in Table 5.1.

5.1.3.1. Characterizing the Population-level Ten-Year Exposure Profile

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of the study population over a ten-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the population-level exposure profiles over two different ten-year periods
- Examine the effect of combining exposure data from multiple shipyards through a multiple imputation approach in characterizing the ten-year exposure profile of the study population at the shipyard of interest

5.1.3.2. Characterizing the Ten-Year Exposure Profile of an SEG

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of an SEG over a ten-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the exposure profiles of one SEG over two different ten-year periods
- Examine the effect of combining exposure data from multiple shipyards through a multiple imputation approach in characterizing the ten-year exposure profile of an SEG at the shipyard of interest

5.1.3.3. Characterizing the One-Year Exposure Profile of an SEG

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of an SEG over a one-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the exposure profiles of one SEG over two different one-year periods
- Examine the effect of combining exposure data from multiple shipyards through a multiple imputation approach in characterizing the one-year exposure profile of an SEG at the shipyard of interest for the shorter time period of one year

5.1.4. Study Population

The study population consisted of shipyard workers from three naval shipyards. In the first section, all workers who met the selection criteria described in Chapter 3 and were employed during the time period of interest were included. In the second and third sections, workers who held the specified job title and met the same selection criteria were included.

Table 5.1 Summary of analyses completed in Chapter 5

Section	Missing Data Pattern	Sub-analysis	Exposure Data Used
3.1 Characterizing the Ten-Year Exposure Profile of a Large Study Population	By random	1980-1990: 90-99% missing 1990-2000: 90-99% missing	Daily
3.2 Characterizing the Ten-Year Exposure Profile of an SEG	By random	1980-1990: 90-99% missing 1990-2000: 90-99% missing	Daily
3.3 Characterizing the One-Year Exposure Profile of an SEG	By random	1980: 90-99% missing 1990: 90-99% missing	Daily

5.2. Methods

5.2.1. Characterizing the Population-level Ten-Year Exposure Profile

If a significant proportion of data are missing for the population of interest, and exposure data of good quality are available for a similar population at a separate facility, then combining exposure data may result in improved estimates of the population-level exposure. This approach is most appropriate when the exposure profiles for each facility are expected to be similar; an attempt should always be made to understand the similarities and differences in exposure patterns between sites. This section compares the population-level exposures over a ten-year period between three separate shipyards and examines the effect of combining exposure data in characterizing the exposure profile of the study population of interest.

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of the study population over a ten-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the population-level exposure profiles over two different ten-year periods
- Examine the effect of combining exposure data from multiple shipyards through an MI approach in characterizing the ten-year exposure profile of the study population at the shipyard of interest

The overall aims of Chapter 3 included understanding common missing data patterns in occupational cohorts and testing the performance of a multiple imputation approach in

estimating population-level exposures when limited measurement data are available. In those exercises, missing exposure data were imputed using available measurements from the same shipyard. However, if the proportion of missing data is high enough, and exposure data of good quality are available for a similar population at another shipyard, the researcher may consider combining exposure data in an effort to better characterize the population-level exposures for the shipyard of interest. This method is most appropriate when the exposure profiles for each shipyard are expected to be similar. Combining data from shipyards with significantly different exposure patterns may result in exposure misclassification of the study population of interest. Therefore, prior to combing exposure data and employing a multiple imputation approach, an attempt should be made to characterize the exposure patterns at each individual shipyard and to understand how they compare.

In this section, the exposure profile of the study population for a given shipyard is estimated using artificially incomplete daily measurement data from that shipyard and supplemental exposure data from at least one of two other shipyards. This exercise reflects the scenario in which exposure data for the facility of interest are limited but more complete exposure records are available for a different location (and are thus treated as surrogate exposure data). The population-level exposure profile for the shipyard of interest is estimated for two separate decades: 1980-1990 and 1990-2000. The exposure data collected at the shipyard of interest were assigned to be randomly missing at varying proportions ranging from 90% to 99% missing. These missing data were then imputed using the remaining exposure data from that shipyard that were not assigned to be missing combined with all exposure data available from at least one of the two other yards. Overall, four separate scenarios were investigated

for each shipyard, which are summarized in Table 5.2. For example, when exposure data were assigned to be missing for NS1, the missing exposure data were imputed 1) using the remaining data from NS1 combined with the exposure data from NS2 (NS1 + NS2); 2) using the remaining data from NS1 combined with the exposure data from NS3 (NS1 + NS3); 3) using the remaining data from NS1 combined with the exposure data from NS2 and NS3 (NS1+ NS2 + NS3); 4) using the available data from NS1 only, for comparison (NS1). The daily radiation exposure measurements were used for these exercises.

Table 5.2 Summary of surrogate data used by scenario

Shipyard with missing data:	NS1	NS2	NS3
Shipyards used to impute missing data			
Scenario 1: two yards	NS1 + NS2	NS2 + NS1	NS3 + NS1
Scenario 2: two yards	NS1 + NS3	NS2 + NS3	NS3 + NS2
Scenario 3: three yards	NS1 + NS2 + NS3	NS2 + NS1 + NS3	NS3 + NS1 + NS2
Scenario 4: one yard	NS1 only	NS2 only	NS3 only

Prior to conducting the analyses, the exposure patterns for each shipyard at each decade of interest were examined and compared. Table 5.3 compares the yearly mean exposure level, using the daily exposure records, between the three shipyards during the 1980-1990 period. Table 5.4 does the same for the 1990-2000 period. As noted in Chapter 3, exposure levels decreased over time. In addition, the mean exposure levels between the yards became more similar over time. Exposure levels during the 1990-2000 timeframe were more similar between yards; thus, it is possible that combining exposure data may be more appropriate, and lead to more accurate estimates of exposure, during the 1990-2000 period.

The observed differences in exposure patterns over time may be partially reflected by differences in the work performed at each yard. Tables 5.5-5.7 examine the top-ranked job

titles for each yard from 1980-1990 based on mean exposure level, number of measurements collected, and number of workers employed, respectively. The job titles with the highest mean exposure level over the 1980-1990 time period varied between the three yards and were often not the same job titles as those that had the greatest number of measurements collected or the greatest number of employed workers. The job titles with the greatest number of measurements collected and the greatest number of employed workers were often similar across yards. Table 5.8 summarizes the work population and exposure data for each yard during the 1980-1990 timeframe. The most significant observed difference in work population across the yards was the race of the worker. Because the percentages differed so greatly between yards, the race variable was removed from the MI analyses. When looking at the summary of exposure data across the yards, one notable difference is the mean exposure level per quarter. The exposure levels fluctuate by quarter, both within a yard and between yards.

Tables 5.9-5.12 summarize the same information but for the 1990-2000 timeframe. Again, may of the same job titles are observed across yards for each ranking; the job titles with the highest mean exposure levels are once again not often similar to those job titles with the greatest number of collected measurements or greatest number of employed workers. The exposure levels are significantly lower during the 1990-2000 time period as compared to the 1980-1990 period. There is also a decrease in the variance level and more similarity across yards when grouping exposures into high, medium, and low bins.

Based on these observations, two separate exercises in which exposure data are estimated for each decade independently is appropriate. Exposure levels in the 1990-2000 time period are lower and appear to be more similar across yards as compared to the 1980-1990 time period; thus, the performance of the MI approach may differ.

Table 5.3. Mean daily exposure level per year: 1980-1990

Year	NS1 Mean (mrem)	NS2 Mean (mrem)	NS3 Mean (mrem)
1980	94.0	306.0	328.9
1981	60.9	210.0	402.3
1982	184.0	207.5	343.6
1983	197.3	150.4	195.2
1984	272.0	197.0	264.3
1985	199.5	227.8	86.5
1986	195.7	400.4	83.3
1987	239.9	210.6	89.4
1988	206.6	205.0	75.2
1989	211.3	219.1	62.5
1990	196.3	187.7	119.6

Table 5.4. Mean daily exposure level per year: 1990-2000

Year	NS1 Mean (mrem)	NS2 Mean (mrem)	NS3 Mean (mrem)
1990	196.3	187.7	119.6
1991	188.7	150.0	149.3
1992	164.3	138.9	150.6
1993	103.0	175.9	101.9
1994	115.7	151.1	47.7
1995	135.5	133.8	79.6
1996	111.4	175.5	81.1
1997	88.4	213.5	160.3
1998	100.8	95.3	122.3
1999	104.9	107.0	138.2
2000	134.9	173.2	126.5

Table 5.5. Top 10 job titles with highest mean daily exposure level over 1980-1990 time period

Yard	NS1	NS2	NS3
Rank	Job Title (mrem)	Job Title (mrem)	Job Title (mrem)
1	Canvas Working (367.0)	Optical Instrument Repairing (457.2)	Loftsman (1737.0)
2	Insulating (215.1)	Heating and Boiler Plant Equipment Mechanic (346.0)	Miscellaneous Marine Maintenance (314.2)
3	Industrial Equipment Mechanic (197.1)	Pipe Covering (253.7)	Pipe Covering (302.8)
4	Boilermaking (180.1)	Industrial Equipment Mechanic (159.3)	Miscellaneous General Equipment Maintenance (261.0)
5	Fabric Working (168.2)	Miscellaneous Marine Maintenance (154.7)	Upholstering (260.3)
6	Machine Tool Operating (156.8)	Insulating (152.8)	Pneumatic Tool Operating (257.9)
7	Miscellaneous Marine Maintenance (153.0)	Metal Forging (149.3)	Lofting (200.3)
8	Pipe Covering (149.3)	Equipment Cleaning (132.8)	Machine Tool Operating (196.0)
9	Electronics Mechanic (142.1)	Shipfitting (129.4)	Test Reactor Cont. (162.0)
10	Miscellaneous Industrial Equipment Maintenance (141.3)	Lofting (119.1)	Miscellaneous Metal Processing (160.5)

Table 5.6. Top 10 job titles with greatest number of measurements collected over 1980-1990 time period

Yard	NS1	NS2	NS3
Rank	Job Title (No. of measurements)	Job Title (No. of measurements)	Job Title (No. of measurements)
1	Pipefitting (8131)	Pipefitting (6990)	Pipefitting (4462)
2	Electrician (4132)	Marine Machinery Mechanic (4887)	Marine Machinery Mechanic (2891)
3	Marine Machinery Mechanic (4076)	Electrician (4478)	Electrician (2633)
4	Miscellaneous General Maintenance and Operations Work (3765)	Miscellaneous Industrial Equipment Maintenance (4206)	Shipfitting (2385)
5	Rigging (3691)	Shipfitting (2887)	Welding (1576)
6	Boilermaking (3230)	Rigging (2480)	Miscellaneous Industrial Equipment Maintenance (1312)
7	Shipfitting (2776)	Welding (2325)	Painting (1211)
8	Welding (2223)	Equipment Cleaning (1845)	Rigging (1164)
9	Sheet Metal Mechanic (1833)	Miscellaneous Metal Work (1802)	Sheet Metal Mechanic (981)
10	Machining (1608)	Insulating (1570)	Miscellaneous Metal Work (927)

Table 5.7. Top 10 job titles with largest number of workers employed 1980-1990 time period

Yard	NS1	NS2	NS3
Rank	Job Title (No. of workers)	Job Title (No. of workers)	Job Title (No. of workers)
1	Pipefitting (733)	Pipefitting (542)	Pipefitting (261)
2	Electrician (460)	Electrician (395)	Electrician (191)
3	Marine Machinery Mechanic (405)	Marine Machinery Mechanic (323)	Shipfitting (172)
4	Rigging (284)	Shipfitting (249)	Marine Machinery Mechanic (162)
5	Boilermaking (266)	Welding (160)	Machining (145)
6	Shipfitting (242)	Rigging (148)	Welding (103)
7	Welding (209)	Industrial Equipment Mechanic (132)	Rigging (84)
8	Electronics Mechanic (184)	Equipment Cleaning (119)	Painting (79)
9	Sheet Metal Mechanic (181)	Machining (118)	Sheet Metal Mechanic (72)
10	Machining (163)	Electronics Mechanic (114)	Insulating (60)

Table 5.8. Summary of work population and exposure data over 1980-1990 time period

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	4220	3341	2024
No. White (%)*	3146 (74.5)	3300 (98.8)	351 (17.3)
No. Black (%)	1007 (23.9)	21 (0.0)	16 (0.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	525 (25.9)
No. other race (%)	67 (1.6)	20 (1.2)	1132 (56.8)
No. HS grad (%)	2883 (68.3)	2384 (71.4)	969 (47.9)
No. terminal occupational program (%)	549 (13.0)	348 (10.4)	378 ((18.7)
No. other education level (%)	788 (81.3)	609 (18.2)	677 (33.4)
No. birth year <1950 (%)	1439 (34.1)	1678 (50.2)	1063 (52.5)
No. birth year >= 1950 (%)	2781 (65.9)	1663 (49.8)	961 (47.5)
Summary of Exposure Data			
Total No. of daily measurements collected	46116	45236	27880
Total No. of 0 mrem measurements (%)	11957 (25.9)	7506 (16.6)	7589 (27.2)
Avg. No. of measurements per worker	2.6	3.0	2.6
All daily measurements			
Mean exposure (mrem)	78.0	75.3	63.0
Median Exposure (mrem)	10.0	16.0	6.0
Peak Exposure (mrem)	8872	1978	1908
Variance	3.43E+04	2.19E+04	3.05E+04
All daily measurements by exposure bin			
% Low (0 to < 5 mrem)	42.7	32.8	46.3
% Medium (5 to <10 mrem)	38.0	45.4	39.5
% High (≥100 mrem)	19.3	21.8	14.2
Mean Exposure by quarter			
Quarter 1 (mrem)	63.5	69.4	34.1
Quarter 2 (mrem)	60.7	85.2	37.0
Quarter 3 (mrem)	102.8	82.6	119.9
Quarter 4 (mrem)	75.4	63.1	33.9

Table 5.9. Top 10 job titles with highest mean daily exposure level over 1990-2000 time period

Yard	NS1	NS2	NS3
Rank	Job Title (mrem)	Job Title (mrem)	Job Title (mrem)
1	Machine Tool Operating (161.4)	Automotive Mechanic (82.3)	Leadburning (48.7)
2	Coppersmithing (62.5)	Metal Forging (75.6)	Lofting (43.7)
3	Insulating (50.4)	Insulating (49.3)	Insulating (25.8)
4	Fabric Working (36.9)	Equipment Cleaning (43.8)	Shipfitting (23.5)
5	Boiler Plant Operating (34.7)	Miscellaneous Woodwork (38.5)	Equipment Cleaning (15.7)
6	Boilermaking (32.4)	Lofting (24.9)	Testing Equipment Operating (15.0)
7	Marine Machinery Mechanic (24.3)	Shipfitting (24.5)	Machine Tool Operating (14.5)
8	Shipfitting (22.9)	Small Engine Mechanic (20.4)	Marine Machinery Mechanic (14.5)
9	Metal Forging (21.7)	Flame/Arc Cutting (18.7)	Fabric Working (13.6)
10	Machining (20.0)	Marine Machinery Mechanic (16.3)	Welding (11.5)

Table 5.10. Top 10 job titles with greatest number of measurements collected over 1990-2000 time period

Yard	NS1	NS2	NS3
Rank	Job Title (No. of measurements)	Job Title (No. of measurements)	Job Title (No. of measurements)
1	Pipefitting (27267)	Pipefitting (22529)	Pipefitting (19464)
2	Rigging (18631)	Marine Machinery Mechanic (20491)	Marine Machinery Mechanic (18430)
3	Marine Machinery Mechanic (14762)	Rigging (16508)	Electrician (8378)
4	Boilermaking (13354)	Miscellaneous Industrial Equipment Maintenance (12338)	Shipfitting (8233)
5	Shipfitting (12413)	Shipfitting (12017)	Rigging (8158)
6	Electrician (12057)	Electrician (10453)	Miscellaneous General Maintenance and Operations Work (6413)
7	Welding (9144)	Welding (8917)	Welding (6057)
8	Miscellaneous General Maintenance and Operations Work (8932)	Painting (7889)	Sheet Metal Mechanic (4085)
9	Sheet Metal Mechanic (5879)	Electronic Measurement Equipment Mechanic (5905)	Painting (3817)
10	Painting (5371)	Miscellaneous Metal Work (4643)	Miscellaneous Industrial Equipment Maintenance (3307)

Table 5.11. Top 10 job titles with largest number of workers employed 1990-2000 time period

Yard	NS1	NS2	NS3
Rank	**Job Title (No. of workers)**	**Job Title (No. of workers)**	**Job Title (No. of workers)**
1	Pipefitting (518)	Pipefitting (339)	Pipefitting (231)
2	Marine Machinery Mechanic (374)	Marine Machinery Mechanic (317)	Marine Machinery Mechanic (206)
3	Electrician (349)	Electrician (239)	Electrician (167)
4	Rigging (299)	Miscellaneous Industrial Equipment Maintenance (193)	Shipfitting (111)
5	Miscellaneous General Maintenance and Operations Work (273)	Shipfitting (176)	Rigging (99)
6	Shipfitting (199)	Welding (118)	Welding (84)
7	Welding (181)	Rigging (113)	Miscellaneous Industrial Equipment Maintenance (75)
8	Boilermaking (167)	Painting (90)	Sheet Metal Mechanic (70)
9	Sheet Metal Mechanic (167)	Electronics Mechanic (89)	Painting (62)
10	Machining (126)	Miscellaneous Metal Work (85)	Miscellaneous Metal Work (55)

Table 5.12. Summary of work population and exposure data over 1990-2000 time period

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	3496	2536	1792
No. White (%)	2460 (70.4)	2504 (98.7)	352 (19.6)
No. Black (%)	948 (27.1)	12 (0.0)	10 (0.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	619 (34.5)
No. other race (%)	88 (2.5)	20 (1.3)	811 (45.9)
No. HS grad (%)	1786 (51.1)	2216 (87.4)	970 (54.1)
No. terminal occupational program (%)	1163 (33.2)	31 (1.2)	149 (8.3)
No. other education level (%)	547 (15.7)	289 (11.4)	673 (37.6)
No. birth year <1950 (%)	904 (25.9)	1135 (44.8)	707 (39.5)
No. birth year >= 1950 (%)	2592 (74.1)	1401 (55.2)	1085 (60.5)
Summary of Exposure Data			
Total No. of daily measurements collected	164479	148760	110896
Total No. of 0 mrem measurements (%)	95954 (58.3)	84000 (56.5)	74351 (67.0)
Avg. No. of measurements per worker	8.3	12.7	11.4
All daily measurements			
Mean exposure (mrem)	16.4	12.3	10.1
Median Exposure (mrem)	0.0	0.0	0.0
Peak Exposure (mrem)	5479	1585	1314
Variance	6.14E+03	2.64E+03	1.93E+03
All daily measurements by exposure bin			
% Low (0 to < 5 mrem)	78.5	79.7	80.8
% Medium (5 to <10 mrem)	17.7	17.0	16.6
% High (≥100 mrem)	3.8	3.3	2.6
Mean Exposure by quarter			
Quarter 1 (mrem)	19.3	13.5	7.9
Quarter 2 (mrem)	16.0	13.7	11.8
Quarter 3 (mrem)	13.9	11.6	12.5
Quarter 4 (mrem)	16.8	10.5	8.8

5.2.2. Characterizing the Ten-Year Exposure Profile of an SEG

Similarly to the section above, if the proportion of missing data for an SEG of interest is great enough, and exposure data of good quality are available for a similar population at a separate facility, then combining exposure data may result in improved estimates of the SEG-level exposure. This section compares the exposure profile of the pipefitting SEG over a ten-year period between three separate shipyards and examines the effect of combining exposure data to estimate the exposure levels of this SEG.

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of an SEG over a ten-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the exposure profiles of one SEG over two different ten-year periods
- Examine the effect of combining exposure data from multiple shipyards through an MI approach in characterizing the ten-year exposure profile of an SEG at the shipyard of interest

Similarly to the previous section, the analyses performed in the section explore the effect of combing data from multiple shipyards in characterizing the ten-year exposure profile of the study population. However, in this scenario, the study population is one specific SEG or job title. In Chapter 4, the ability to characterize the exposure profile of an SEG using limited measurement data from the study population of interest was explored. However, if available exposure data for this SEG population is insufficient, researchers and/or industrial hygienists

may have the opportunity to supplement the exposure data with measurements collected on the same SEG or job title, during the same timeframe, but at different facilities.

Examining one specific SEG or job title allows for easier comparisons between yards, as the data are not influenced by differences between job titles. However, it is still important to understand the exposure patterns within each yard and how they compare. For these analyses, the pipefitting SEG is used. As with the previous section, two separate decades will be examined: 1980-1990 and 1990-2000. The exposure profile of the pipefitting SEG for a given shipyard is estimated using artificially incomplete daily measurement data from that shipyard and supplemental exposure data from at least one of two other shipyards. The exposure data collected at the shipyard of interest were assigned to be randomly missing at varying proportions ranging from 50% to 99% missing; 50% missing was added due to the smaller sample size of working with one SEG. These missing data were then imputed using the remaining exposure data from that shipyard that were assigned as not missing combined with all exposure data available from at least one of the two other yards. The same four separate scenarios as were described in the previous section were repeated here. The daily radiation exposure measurements were used for these exercises.

Prior to conducting the analyses, the exposure patterns of the pipefitting SEG for each shipyard were examined and compared. Table 5.13 compares the pipefitting yearly mean exposure level, using the daily exposure records, between the three shipyards during the time period of interest: 1980-2000. As was observed when examining the entire study population, the exposure levels decreased with time and mean exposure levels between the yards became

more similar over time. Exposure levels during the 1990-2000 timeframe were more similar between yards. The observed differences in exposure levels between yards were further examined by the number of collected measurements and number of employed workers per year. Tables 5.14 and 5.15 summarize the work population and exposure data for each yard during the 1980-1990 and 1990-2000 timeframes, respectively. Again, a significant difference in the race variable across yards was observed and thus the variable was removed from the MI analyses. The exposure levels are significantly lower during the 1990-2000 time period as compared to the 1980-1990 time period. There is also a decrease in the variance level and more similarity across yards when grouping exposures into high, medium, and low bins.

Based on these observations, two separate exercises in which exposure data are estimated for each decade independently is appropriate. Even within the same SEG, exposure levels can vary over time, both within and between yards; this variation may be greater in some time periods as compared to others. Thus, the performance of the MI approach may differ between time periods.

Table 5.13. Comparison of mean daily exposure levels, number of measurements collected, and number of workers employed by shipyard for pipefitting SEG

Yard	NS1			NS2			NS3		
Year	Mean (mrem)	No. of meas.	No. of workers	Mean (mrem)	No. of meas.	No. of workers	Mean (mrem)	No. of meas.	No. of workers
1980	68.8	49	44	103.0	342	159	391.7	55	55
1981	36.1	48	16	71.5	298	25	573.3	56	10
1982	121.3	345	284	78.1	523	101	417.0	165	110
1983	134.5	354	79	42.4	738	61	227.5	146	5
1984	63.6	1146	56	61.4	726	20	62.3	517	13
1985	38.5	980	19	67.1	744	27	24.5	529	11
1986	32.9	992	49	76.3	690	24	19.1	569	33
1987	63.4	923	11	48.5	636	41	29.3	582	10
1988	56.1	1034	65	65.5	713	35	17.5	585	12
1989	52.4	1060	69	58.9	819	58	19.0	651	12
1990	43.1	1200	75	53.2	761	12	31.0	607	7
1991	52.0	1288	77	44.2	767	32	42.8	504	11
1992	38.8	1538	47	37.1	647	25	46.6	542	17
1993	25.7	1316	15	70.0	631	58	34.8	502	4
1994	43.2	864	17	27.9	489	22	15.2	410	3
1995	34.6	760	15	37.7	334	2	33.7	336	3
1996	31.3	941	7	47.8	284	0	19.7	323	3
1997	20.4	708	3	5.9	2664	2	35.1	360	26
1998	29.9	617	1	1.0	6686	2	12.2	1106	7
1999	1.3	11635	10	0.9	5889	0	2.0	9558	3
2000	3.5	6400	7	2.6	3377	8	2.9	5216	3

Table 5.14. Summary of work population and exposure data over 1980-1990 time period: pipefitting SEG

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	767	563	278
No. White (%)	572 (74.6)	559 (99.3)	30 (10.8)
No. Black (%)	185 (24.1)	2 (<1.0)	2 (<1.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	98 (35.3)
No. other race (%)	10 (1.3)	2 (<1.0)	148 (53.2)
No. high school grad (%)	574 (74.8)	410 (72.8)	135 (48.6)
No. terminal occupational program (%)	82 (10.7)	56 (9.9)	50 (18.0)
No. other education level	111 (14.5)	97 (17.3	93 (33.4)
No. birth year <1950 (%)	266 (34.7)	265 (47.1)	141 (50.7)
No. birth year >= 1950 (%)	501 (65.3)	298 (52.9)	137 (49.3)
Summary of Exposure Data			
Total No. of daily measurements collected	8131	6990	4462
Total No. of 0 mrem measurements (%)	2119 (26.1)	856 (12.2)	1081 (24.2)
Avg. No. of measurements per worker	2.7	3.1	2.8
All daily measurements			
Mean exposure (mrem)	56.8	63.3	60.5
Median Exposure (mrem)	8.0	21.5	7.0
Peak Exposure (mrem)	1782	1872	1896
Variance	1.46E+04	1.27E+04	3.35E+04
All daily measurements by exposure bin			
Low (0 to < 5 mrem)	44.0	25.5	44.2
Medium (5 to <10 mrem)	39.1	55.9	43.4
High (≥100 mrem)	16.9	18.6	12.4
Mean Exposure by quarter			
Quarter 1 (mrem)	46.0	65.0	28.0
Quarter 2 (mrem)	51.0	61.1	30.0
Quarter 3 (mrem)	72.0	70.9	124.4
Quarter 4 (mrem)	52.7	55.1	34.3

Table 5.15. Summary of work population and exposure data over 1990-2000 time period: pipefitting SEG

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	576	372	240
No. White (%)	413 (71.7)	372 (1.0)	38 (15.8)
No. Black (%)	152 (26.4)	0 (0.0)	2 (<1.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	107 (44.6)
No. other race (%)	11 (1.9)	0 (0.0)	93 (38.8)
No. high school grad (%)	318 (55.2)	330 (88.7)	121 (50.4)
No. terminal occupational program (%)	176 (30.6)	3 (<1.0)	21 (8.8)
No. other education level	82 (14.2)	39 (10.5)	98 (40.8)
No. birth year <1950 (%)	145 (25.2)	160 (43.0)	96 (40.0)
No. birth year >= 1950 (%)	431 (74.8)	212 (57.0)	144 (60.0)
Summary of Exposure Data			
Total No. of daily measurements collected	27267	22529	19464
Total No. of 0 mrem measurements (%)	16172 (59.3)	12612 (56.0)	13052 (67.1)
Avg. No. of measurements per worker	8.4	14.8	15.1
All daily measurements			
Mean exposure (mrem)	13.8	9.7	8.6
Median Exposure (mrem)	0.0	0.0	0.0
Peak Exposure (mrem)	1494	1098	557
Variance	3.39E+03	1.87E+03	1.08E+03
All daily measurements by exposure bin			
Low (0 to < 5 mrem)	79.3	81.9	81.5
Medium (5 to <10 mrem)	17.0	15.7	16.1
High (≥100 mrem)	3.7	2.4	2.4
Mean Exposure by quarter			
Quarter 1 (mrem)	14.7	10.6	6.7
Quarter 2 (mrem)	13.4	10.2	10.8
Quarter 3 (mrem)	11.5	9.7	8.8
Quarter 4 (mrem)	16.9	8.6	8.7

5.2.3. Characterizing the One-Year Exposure Profile of an SEG

Exposure data for an SEG may also be limited even within the shorter timeframe of one year. In this case, supplemental exposure data from separate shipyards may be less varied, given the shorter time period; this may result in more accurate estimates of the true exposure profile than when compared to estimates obtained using a ten-year period of data. This section compares the exposure profile of the pipefitting SEG over a one-year period between three separate shipyards and examines the effect of combining exposure data to estimate the exposure levels of this SEG for the shorter time period of one year.

In this section, surrogate exposure data from additional shipyards are used to characterize the exposures of an SEG over a one-year period at the shipyard of interest. The analyses in this section address the following research objectives:

- Compare between shipyards the exposure profiles of one SEG over two different one-year periods
- Examine the effect of combining exposure data from multiple shipyards through an MI approach in characterizing the one-year exposure profile of an SEG at the shipyard of interest for the shorter time period of one year

In this section, the one-year exposure profiles of two separate SEGs are estimated using a combination of exposure data from the shipyard of interest and at least one of two other yards. This exercise reflects the scenario in which an industrial hygienist and/or researcher wishes to characterize the exposure profile for a particular population of workers, all of whom are assigned to the same SEG, during a short timeframe; however, exposure data for

this specific work population is limited. The reasons for limited availability of exposure data for an SEG have been discussed previously and include financial constraints on the number of samples that can be collected and poor data retention policies. Regardless, more complete exposure data may be available for the same SEG, and during the same short timeframe, but at another facility. Combining the available exposure data may allow for an improved characterization of the exposure profile of the SEG at the shipyard of interest.

Unlike the previous section, which characterized the exposure profile of an SEG over a 10-year period, this section focuses on a shorter time period: one year. For these analyses, the pipefitting SEG is again used. Two one-year periods are examined: the year 1980 and the year 1990. The exposure profile of the pipefitting SEG for a given shipyard is estimated using artificially incomplete daily measurement data from that shipyard and supplemental exposure data from at least one of two other shipyards. The exposure data collected at the shipyard of interest were assigned to be randomly missing at varying proportions ranging from 50% to 99% missing; 50% missing was added due to the smaller sample size of working with one SEG. These missing data were then imputed using the remaining exposure data from that shipyard that were assigned as not missing combined with all exposure data available from at least one of the two other yards. The same four separate scenarios as were described in the previous sections were repeated here. The daily radiation exposure measurements were used for these exercises.

Tables 5.16 and 5.17 summarize the work population and exposure data of the pipefitting SEG for the year 1980 and 1990, respectively. The overall patterns for the work population

are similar to what was observed when looking at a ten-year period, with the exception of the unknown race data for NS3 in 1980. However, the exposure levels between yards were quite varied for the year 1980. One of the most striking differences is the comparison of the mean exposure by quarter. The exposure levels between the yards for the year 1990 were observed to vary less.

Based on these observations, two separate exercises in which exposure data are estimated independently for two different years is appropriate. Even within the same SEG, exposure levels can vary over time, both within and between yards; this variation may be greater in during one year as compared to another. In addition, unlike when looking at a ten-year period, differences between yards may be more pronounced during a shorter timeframe. Thus, the performance of the MI approach may differ based on the year of the collected data.

Table 5.16. Summary of work population and exposure data during the year 1980: pipefitting SEG

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	44	159	55
No. White (%)	39 (88.6)	159 (100.0)	0 (0.0)
No. Black (%)	5 (11.4)	0 (0.0)	0 (0.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	0 (0.0)
No. Unknown (%)	0 (0.0)	0 (0.0)	55 (100.0)
No. high school grad (%)	10 (22.8)	69 (43.4)	12 (21.8)
No. terminal occupational program (%)	13 (29.5)	44 (27.7)	27 (49.1)
No. birth year <1950 (%)	33 (75.0)	116 (73.0)	48 (87.3)
No. birth year >= 1950 (%)	11 (25.0)	43 (27.0)	7 (12.7)
Summary of Exposure Data			
Total No. of daily measurements collected	49	342	55
Total No. of 0 mrem measurements (%)	19 (38.8)	37 (10.8)	11 (20.0)
Avg. No. of measurements per worker	1.1	2.3	1.0
All daily measurements			
Mean exposure (mrem)	68.8	103.0	391.7
Median Exposure (mrem)	4.0	34.5	136.0
Peak Exposure (mrem)	677	1026	1611
Variance	2.24E+04	2.86E+04	2.47E+05
All daily measurements by exposure bin			
Low (0 to < 5 mrem)	51.0	24.6	20.0
Medium (5 to <10 mrem)	28.6	46.2	23.6
High (≥100 mrem)	20.4	29.2	56.4
Mean Exposure by quarter			
Quarter 1 (mrem)	0.0	59.5	--
Quarter 2 (mrem)	0.0	56.9	--
Quarter 3 (mrem)	76.6	166.3	391.7
Quarter 4 (mrem)	--	71.8	--

Table 5.17. Summary of work population and exposure data during the year 1990: pipefitting SEG

	NS1	NS2	NS3
Summary of Work Population			
Total No. of workers	377	221	160
No. White (%)	272 (72.1)	221 (100.0)	14 (8.8)
No. Black (%)	99 (26.3)	0 (0.0)	0 (0.0)
No. Japanese (%)	0 (0.0)	0 (0.0)	79 (49.4)
No. high school grad (%)	216 (57.3)	198 (89.6)	85 (53.1)
No. terminal occupational program (%)	105 (27.9)	1 (<1.0)	12 (7.5)
No. birth year <1950 (%)	98 (26.0)	83 (37.6)	69 (43.1)
No. birth year >= 1950 (%)	279 (74.0)	138 (19.5)	91 (56.9)
Summary of Exposure Data			
Total No. of daily measurements collected	1200	761	607
Total No. of 0 mrem measurements (%)	434 (36.2)	99 (13.0)	195 (32.1)
Avg. No. of measurements per worker	3.3	3.5	3.9
All daily measurements			
Mean exposure (mrem)	43.1	53.2	31.0
Median Exposure (mrem)	3.0	22.0	3.0
Peak Exposure (mrem)	1322	1049	467
Variance	1.10E+04	9.46E+03	4.45E+03
All daily measurements by exposure bin			
Low (0 to < 5 mrem)	52.9	26.4	53.2
Medium (5 to <10 mrem)	34.3	59.0	37.2
High (≥100 mrem)	12.8	14.6	9.6
Mean Exposure by quarter			
Quarter 1 (mrem)	47.3	36.4	7.7
Quarter 2 (mrem)	39.9	68.8	41.6
Quarter 3 (mrem)	53.0	51.6	19.3
Quarter 4 (mrem)	32.1	57.0	58.4

5.3. Results

5.3.1. Characterizing the Population-level Ten-Year Exposure Profile

1980-1990 time period

Estimated Population Mean

When estimating the exposures for NS1, the direction of the bias varied based on whether the data from NS1 were combined with NS2 or NS3 (Table 5.18-5.20). Combining data from all three yards resulted in improved accuracy. Similar patterns were observed when estimating the exposures for NS2 and NS3 as well. When estimating exposures for NS2, using data from NS2 combined with NS1 produced some of the least biased estimates; however, combining data from NS2 with data from NS3 produced some of the most biased results. Using only the remaining available data from the shipyard of interest resulted in some of the least accurate estimates of the mean and the widest confidence intervals.

Estimated Population Median

Combining data from all three yards generally produced estimates of the median that were neither the least nor most biased (Tables 5.18-5.20). When estimating the median for NS1, combining data from all three yards resulted in estimates of the median that were more accurate compared to using either combination of two yards.

Estimated Imputation Variance

Combining data from all three yards often resulted in estimates of the variance that were neither the least or most biased (Tables 5.18-5.20). When estimating the variance for the data

at NS2, most of the imputations variances overestimated the true variance. Using only the remaining available data from the shipyard of interest resulted in overestimates of the true variance as the percentage of missing data increased.

1990-2000 time period

Estimated Population Mean

For each of the three yards, all trials resulted in overestimations of the true mean (Tables 5.21-5.23). For NS1, combining data from all three shipyards resulted in estimates of the mean that were the least biased; for NS2 and NS3, the estimates produced by combining data from all three yards generally were not the most biased.

Estimated Population Median

Nearly all trials resulted in estimates of the median that were the same as the true median, with the exception of some of the estimates for NS1 (Tables 5.21-5.23).

Estimated Imputation Variance

Many of the trials resulted in estimates of the variance that were overestimates of the true variance **(**Tables 5.21-5.23). Exceptions were observed when estimating the variance for NS1.

Table 5.18. Performance of MI by percent missing from NS1: 1980-1990 time period (mean = 78.0 mrem, median = 10.0 mrem, SE = 185.2)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
90.0	-2.0	(0, 378)	5.6	2.38E+04	0.9	154.2
95.0	-0.2	(0, 391)	5.8	2.56E+04	0.3	160.0
99.0	7.2	(0, 461)	5.2	3.68E+04	0.3	191.8
NS1 + NS3						
90.0	-12.6	(0, 408)	-3.0	3.06E+04	1.7	174.9
95.0	-15.4	(0, 395)	-3.2	2.88E+04	2.6	169.7
99.0	-13.8	(0, 408)	-4.0	3.08E+04	3.5	175.6
NS1 + NS2 + NS3						
90.0	-7.2	(0, 383)	1.8	2.55E+04	0.2	159.6
95.0	-7.8	(0, 380)	1.8	2.50E+04	0.1	158.1
99.0	-8.9	(0, 376)	1.2	2.46E+04	0.5	156.9
NS1 only						
90.0	0.3	(0, 426)	0.4	3.16E+04	5.8	177.7
95.0	7.2	(0, 472)	0.2	3.90E+04	8.2	197.5
99.0	38.4	(0, 625)	2.2	6.73E+04	73.7	259.6

Table 5.19. Performance of MI by percent missing from NS2: 1980-1990 time period (mean = 75.3 mrem, median = 16.0 mrem, SE = 148.0)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2 + NS1						
90.0	0.8	(0, 435)	-6.2	3.37E+04	0.3	183.6
95.0	0.8	(0, 430)	-6.6	3.27E+04	0.6	180.8
99.0	1.0	(0, 435)	-7.0	3.36E+04	2.1	183.2
NS2 + NS3						
90.0	-10.4	(0, 400)	-8.8	2.92E+04	2.4	171.0
95.0	-9.8	(0, 409)	-9.0	3.08E+04	1.2	175.5
99.0	-9.9	(0, 417)	-10.0	3.23E+04	1.9	179.7
NS2 + NS1 + NS3						
90.0	-3.8	(0, 419)	-8.0	3.15E+04	0.3	177.5
95.0	-4.5	(0, 418)	-8.0	3.14E+04	0.0	177.1
99.0	-4.3	(0, 419)	-8.0	3.16E+04	0.2	177.7
NS2 only						
90.0	-4.6	(0, 354)	-1.2	2.09E+04	3.1	144.7
95.0	0.0	(0, 376)	1.2	2.36E+04	9.0	153.7
99.0	5.6	(0, 390)	6.6	2.49E+04	42.0	158.0

Table 5.20. Performance of MI by percent missing from NS3: 1980-1990 time period (mean = 63.0 mrem, median = 6.0 mrem, SE = 174.8)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
90.0	12.0	(0, 430)	2.8	3.29E+04	0.6	181.3
95.0	13.1	(0, 432)	3.0	3.31E+04	0.7	182.0
99.0	12.7	(0, 436)	3.0	3.38E+04	0.2	183.8
NS3 + NS2						
90.0	10.6	(0, 363)	9.0	2.18E+04	0.0	147.7
95.0	11.3	(0, 363)	9.6	2.17E+04	0.4	147.3
99.0	12.5	(0, 367)	10.0	2.22E+04	0.8	148.9
NS3 + NS1 + NS2						
90.0	12.7	(0, 401)	6.4	2.77E+04	0.1	166.4
95.0	13.0	(0, 402)	6.8	2.78E+04	0.1	166.6
99.0	13.1	(0, 401)	7.0	2.76E+04	0.0	166.0
NS3 only						
90.0	0.5	(0, 409)	0.2	3.11E+04	14.0	176.3
95.0	3.4	(0, 458)	-0.2	3.99E+04	6.6	199.9
99.0	28.6	(0, 605)	0.0	6.79E+04	793.6	262.4

Table 5.21. Performance of MI by percent missing from NS1: 1990-2000 time period (mean = 16.4 mrem, median = 0.0 mrem, SE = 78.3)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
90.0	4.1	(0, 164)	0.0	5.36E+03	0.1	73.2
95.0	6.6	(0, 170)	1.0	5.65E+03	0.4	75.2
99.0	7.7	(0, 171)	1.0	5.65E+03	0.5	75.2
NS1 + NS3						
90.0	2.6	(0, 155)	0.0	4.84E+03	0.3	69.5
95.0	2.0	(0, 146)	0.0	4.26E+03	0.3	65.2
99.0	2.3	(0, 142)	0.0	4.00E+03	0.5	63.3
NS1 + NS2 + NS3						
90.0	0.8	(0, 139)	0.0	3.91E+03	0.1	62.5
95.0	1.7	(0, 145)	0.0	4.23E+03	0.2	65.0
99.0	0.7	(0, 135)	0.0	3.68E+03	0.2	60.6
NS1 only						
90.0	3.8	(0, 192)	0.0	7.70E+03	2.0	87.7
95.0	2.6	(0, 181)	0.0	6.89E+03	0.9	83.0
99.0	4.6	(0, 235)	0.0	1.20E+04	12.0	109.6

Table 5.22. Performance of MI by percent missing from NS2: 1990-2000 time period (mean = 12.3 mrem, median = 0.0 mrem, SE = 51.4)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2 + NS1						
90.0	5.7	(0, 172)	0.0	6.22E+03	0.1	78.9
95.0	6.3	(0, 177)	0.0	6.61E+03	0.1	81.3
99.0	7.1	(0, 188)	0.0	7.40E+03	0.1	86.0
NS2 + NS3						
90.0	3.8	(0, 128)	0.0	3.29E+03	0.1	57.3
95.0	4.3	(0, 129)	0.0	3.34E+03	0.3	57.8
99.0	4.7	(0, 131)	0.0	3.38E+03	0.1	58.1
NS2 + NS1 + NS3						
90.0	6.3	(0, 177)	0.0	6.61E+03	0.1	81.3
95.0	7.1	(0, 188)	0.0	7.40E+03	0.1	86.0
99.0	4.2	(0, 158)	0.0	5.22E+03	0.0	72.3
NS2 only						
90.0	4.6	(0, 140)	0.0	3.96E+03	0.2	62.9
95.0	5.0	(0, 143)	0.0	4.17E+03	1.7	64.6
99.0	1.3	(0, 119)	0.0	2.91E+03	7.5	54.0

Table 5.23. Performance of MI by percent missing from NS3: 1990-2000 time period (mean = 10.1 mrem, median = 0.0 mrem, SE = 43.9)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
90.0	7.4	(0, 171)	0.0	6.20E+03	0.05	78.7
95.0	8.3	(0, 179)	0.0	6.71E+03	0.1	81.9
99.0	9.4	(0, 186)	0.0	7.24E+03	0.2	85.1
NS3 + NS2						
90.0	6.2	(0, 132)	0.0	3.49E+03	0.2	59.1
95.0	6.4	(0, 132)	0.0	3.52E+03	0.2	59.4
99.0	7.4	(0, 137)	0.0	3.76E+03	0.2	61.3
NS3 + NS1 + NS2						
90.0	6.2	(0, 154)	0.0	4.93E+03	0.03	70.2
95.0	6.3	(0, 155)	0.0	5.04E+03	0.05	71.0
99.0	6.7	(0, 159)	0.0	5.27E+03	0.02	72.6
NS3 only						
90.0	3.2	(0, 105)	0.0	2.22E+03	0.6	47.2
95.0	2.0	(0, 99)	0.0	1.98E+03	0.3	44.5
99.0	1.3	(0, 103)	0.0	2.19E+03	1.6	46.8

5.3.2. Characterizing the Ten-Year Exposure Profile of an SEG

1980-1990 time period

Estimated SEG Mean

When estimating the MI mean for the shipyard of interest, the direction of the bias varied based on which shipyards' data were combined (Tables 5.24-5.26). Combining data from all three yards often resulted in estimates that were neither the least or most biased. In some cases, combining data from all three yards improved the accuracy of the MI mean compared to using just two of the yards.

Estimated SEG Median

When estimating the MI median for the shipyard of interest, the direction of the bias varied based on which shipyards' data were combined (Tables 5.24-5.26). Combining data from all

three yards sometimes improved the accuracy of the estimates compared to using data from only two yards.

Estimated Imputation Variance

Combining data from all three yards generally produced estimates of the variance that were neither the least nor most biased (Tables 5.24-5.26). When the percentage of missing data was set to 99% missing, using only the available data from the shipyard of interest resulted in estimates of the variance that were the least accurate.

1990-2000 time period

Estimated SEG Mean

Most of the trials within each shipyard resulted in overestimates of the true mean (Tables 5.27-5.29). Combining data from all three yards generally produced estimates that were more biased when estimating the MI mean for NS1 and less biased when estimating the MI mean for NS3.

Estimated SEG Median

Many of the trials estimated the true median exactly (Tables 5.27-5.29). In most of the remaining trials, the estimated median was a slight overestimation of the true median.

Estimated Imputation Variance

Combining data from all three yards generally produced estimates of the variance that were neither the least nor most biased (Tables 5.27-5.29). As the percentage of missing data

increased, the estimates of the variance obtained from using only the available data from the shipyard of interest generally became less accurate.

Table 5.24. Performance of MI by percent missing from NS1, pipefitting SEG: 1980-1990 time period (mean = 56.8 mrem, median = 8.0 mrem, SE = 121.0)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
50.0	5.5	(0, 298)	8.8	1.46E+04	0.6	120.7
90.0	7.7	(0, 293)	11.8	1.36E+04	1.2	116.6
95.0	7.3	(0, 289)	13.0	1.32E+04	1.3	115.0
99.0	6.4	(0, 282)	13.6	1.25E+04	1.8	111.7
NS1 + NS3						
50.0	4.7	(0, 376)	0.2	2.58E+04	1.0	160.7
90.0	9.6	(0, 422)	0.6	3.29E+04	132.8	181.7
95.0	10.8	(0, 445)	-0.4	3.69E+04	141.1	192.6
99.0	17.1	(0, 484)	0.0	4.37E+04	129.9	209.4
NS1 + NS2 + NS3						
50.0	5.4	(0, 336)	5.8	1.96E+04	0.2	140.1
90.0	6.4	(0, 349)	6.4	2.13E+04	6.0	145.9
95.0	8.0	(0, 363)	7.0	2.33E+04	0.1	152.6
99.0	8.3	(0, 363)	7.6	2.31E+04	2.5	152.0
NS1 only						
50.0	1.5	(0, 303)	1.0	1.57E+04	2.8	125.3
90.0	-7.8	(0, 236)	-2.6	9.08E+03	25.8	95.4
95.0	-5.3	(0, 239)	-1.4	9.07E+03	78.9	95.7
99.0	-20.8	(0, 180)	-4.8	5.31E+03	122.1	73.9

Table 5.25. Performance of MI by percent missing from NS2, pipefitting SEG: 1980-1990 time period (mean = 63.3 mrem, median = 21.5 mrem, SE = 112.7)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2 + NS1						
50.0	-5.8	(0, 283)	-9.5	1.33E+04	0.2	115.1
90.0	-6.5	(0, 291)	-12.5	1.44E+04	1.3	119.9
95.0	-8.0	(0, 288)	-13.9	1.41E+04	1.2	118.9
99.0	-8.1	(0, 290)	-14.1	1.44E+04	0.2	120.1
NS2 + NS3						
50.0	-0.9	(0, 365)	-8.5	2.39E+04	10.2	154.7
90.0	8.3	(0, 442)	-9.7	3.57E+04	82.4	189.3
95.0	5.3	(0, 440)	-11.9	3.60E+04	64.6	189.9
99.0	9.6	(0, 474)	-12.7	4.17E+04	158.9	204.7
NS2 + NS1 + NS3						
50.0	-3.4	(0, 335)	-10.5	1.97E+04	0.6	140.4
90.0	-4.3	(0, 345)	-12.7	2.13E+04	0.3	146.0
95.0	-5.6	(0, 340)	-13.5	2.08E+04	0.3	144.2
99.0	-4.1	(0, 352)	-13.5	2.24E+04	1.9	149.7
NS2 only						
50.0	-0.9	(0, 278)	-0.5	1.22E+04	3.7	110.4
90.0	3.8	(0, 206)	1.3	1.49E+04	13.7	122.0
95.0	-10.5	(0, 222)	-3.5	7.51E+03	15.5	86.8
99.0	-11.0	(0, 220)	-8.3	7.17E+03	143.2	85.7

Table 5.26. Performance of MI by percent missing from NS3, pipefitting SEG: 1980-1990 time period (mean = 60.5 mrem, median = 7.0 mrem, SE = 183.0)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
50.0	-4.0	(0, 320)	0.2	1.81E+04	0.1	134.6
90.0	-6.9	(0, 288)	0.0	1.43E+04	0.5	119.6
95.0	-5.9	(0, 288)	0.2	1.42E+04	0.2	119.1
99.0	-5.1	(0, 290)	0.6	1.44E+04	1.0	119.9
NS3 + NS2						
50.0	1.0	(0, 315)	10.0	1.68E+04	0.3	129.8
90.0	0.2	(0, 286)	11.8	1.32E+04	1.7	115.0
95.0	0.4	(0, 278)	12.6	1.23E+04	1.2	111.1
99.0	1.8	(0, 282)	14.0	1.26E+04	1.2	112.3
NS3 + NS1 + NS2						
50.0	-1.2	(0, 308)	6.0	1.61E+04	0.2	126.9
90.0	-2.2	(0, 289)	6.2	1.39E+04	0.7	118.0
95.0	-2.1	(0, 286)	6.6	1.35E+04	0.4	116.2
99.0	-1.0	(0, 290)	7.0	1.39E+04	0.7	118.0
NS3 only						
50.0	3.8	(0, 422)	0.4	3.35E+04	11.9	183.0
90.0	-10.0	(0, 351)	-1.8	2.34E+04	160.2	153.5
95.0	-17.7	(0, 368)	-4.2	2.73E+04	254.1	166.0
99.0	87.0	(0, 754)	2.0	9.26E+04	2674.0	309.6

Table 5.27. Performance of MI by percent missing from NS1, pipefitting SEG: 1990-2000 time period (mean = 13.8 mrem, median = 0.0 mrem, SE = 58.2)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
50.0	0.7	(0, 124)	0.0	3.17E+03	0.2	56.3
90.0	7.0	(0, 156)	1.0	4.81E+03	1.6	69.4
95.0	7.9	(0, 159)	1.0	4.93E+03	1.1	70.2
99.0	4.9	(0, 143)	1.0	4.02E+03	2.2	63.5
NS1 + NS3						
50.0	0.5	(0, 115)	0.0	2.67E+03	0.0	51.6
90.0	4.8	(0, 130)	0.0	3.25E+03	1.5	57.1
95.0	4.5	(0, 121)	0.0	2.76E+03	0.3	52.5
99.0	5.5	(0, 123)	0.0	2.85E+03	1.2	53.4
NS1 + NS2 + NS3						
50.0	-0.5	(0, 336)	0.0	2.52E+03	0.1	50.2
90.0	3.1	(0, 349)	0.0	3.17E+03	0.5	56.3
95.0	3.6	(0, 363)	0.0	3.07E+03	0.5	55.4
99.0	4.5	(0, 363)	0.0	3.23E+03	0.7	56.9
NS1 only						
50.0	2.9	(0, 303)	0.0	3.55E+03	0.3	59.5
90.0	2.5	(0, 236)	0.0	4.40E+03	4.3	66.4
95.0	4.4	(0, 239)	0.0	4.54E+03	7.6	67.5
99.0	12.9	(0, 180)	0.0	9.26E+03	20.5	96.4

Table 5.28. Performance of MI by percent missing from NS2, pipefitting SEG: 1990-2000 time period (mean = 9.7 mrem, median = 0.0 mrem, SE = 43.3)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2 + NS1						
50.0	4.1	(0, 124)	0.0	3.18E+03	0.0	56.4
90.0	5.3	(0, 128)	0.0	3.37E+03	0.7	58.1
95.0	7.0	(0, 140)	0.0	3.99E+03	0.3	63.1
99.0	7.6	(0, 146)	0.0	4.33E+03	1.4	65.8
NS2 + NS3						
50.0	2.1	(0, 96)	0.0	1.87E+03	0.1	43.3
90.0	3.1	(0, 92)	0.0	1.65E+03	1.4	40.7
95.0	3.3	(0, 93)	0.0	1.69E+03	1.0	41.2
99.0	5.0	(0, 100)	0.0	1.93E+03	0.6	43.9
NS2 + NS1 + NS3						
50.0	3.2	(0, 115)	0.0	2.73E+03	0.1	52.3
90.0	4.7	(0, 122)	0.0	2.52E+03	0.4	50.3
95.0	4.7	(0, 122)	0.0	3.02E+03	0.3	55.0
99.0	5.0	(0, 122)	0.0	3.02E+03	1.1	55.0
NS2 only						
50.0	5.4	(0, 122)	0.8	2.97E+03	5.1	54.6
90.0	3.8	(0, 96)	0.6	1.79E+03	13.2	42.4
95.0	0.2	(0, 88)	0.0	1.60E+03	6.6	40.2
99.0	-4.2	(0, 43)	0.0	3.65E+02	1.2	19.1

Table 5.29. Performance of MI by percent missing from NS3, pipefitting SEG: 1990-2000 time period (mean = 8.6 mrem, median = 0.0 mrem, SE = 32.9)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
50.0	4.4	(0, 116)	0.0	2.81E+03	0.1	53.0
90.0	6.3	(0, 132)	0.0	3.58E+03	0.3	59.9
95.0	7.0	(0, 135)	0.0	3.76E+03	0.1	61.3
99.0	7.6	(0, 138)	0.0	3.90E+03	2.1	62.5
NS3 + NS2						
50.0	2.0	(0, 95)	0.0	1.88E+03	0.1	43.4
90.0	5.9	(0, 119)	0.0	2.88E+03	2.2	53.7
95.0	4.4	(0, 109)	0.0	2.44E+03	1.3	49.4
99.0	4.4	(0, 109)	0.2	2.41E+03	3.5	49.1
NS3 + NS1 + NS2						
50.0	3.1	(0, 108)	0.0	2.41E+03	0.0	49.1
90.0	3.4	(0, 110)	0.0	2.53E+03	0.1	50.3
95.0	3.9	(0, 112)	0.0	2.61E+03	0.3	51.1
99.0	3.9	(0, 112)	0.0	2.62E+03	0.3	51.2
NS3 only						
50.0	4.3	(0, 94)	0.0	1.71E+03	0.3	41.4
90.0	2.6	(0, 89)	0.0	1.57E+03	4.0	39.7
95.0	1.2	(0, 79)	0.0	1.24E+03	5.1	35.3
99.0	5.6	(0, 108)	0.0	2.30E+03	8.3	48.1

5.3.3. Characterizing the One-Year Exposure Profile of an SEG

1980 time period

Estimated SEG Mean

Combining data from all three yards generally produced MI mean estimates that were neither the least nor most biased (Tables 5.30-5.32). The direction and magnitude of the estimates when only two yards were combined varied depending on which yards were used. At higher percentages (90% or 95%), there were too few measurements for the MI procedure to work properly for NS1 and NS3.

Estimated SEG Median

Similar to what was observed when estimating the MI mean, combining data from all three yards generally produced MI median estimates that were neither the least nor most biased (Tables 5.30-5.32). The direction and magnitude of the estimates when only two yards were combined varied depending on which yards were used.

Estimated Imputation Variance

When estimating the variance for NS1 and NS2, a majority of the trials produced imputation variances that were overestimates of the true variance (Tables 5.30-5.32). The opposite was observed when estimating the variance for NS3.

1990 time period

Estimated SEG Mean

Combining data from all three yards generally produced MI mean estimates that were neither the least nor most biased (Tables 5.33-5.35). The direction and magnitude of the estimates when only two yards were combined varied depending on which yards were used. This was similar to what was observed during the 1980 time period.

Estimated SEG Median

When estimating the MI median for NS1 and NS3, a majority of the trials produced overestimates of the true median (Tables 5.33-5.35). The opposite was observed when estimating the MI median for NS2.

Estimated Imputation Variance

Combining data from all three yards generally produced estimates of the variance that were neither the least nor most biased (Tables 5.33-5.35). As the percentage of missing data increased, the estimates of the variance obtained from using only the available data from the shipyard of interest generally became less accurate.

Table 5.30. Performance of MI by percent missing from NS1, pipefitting SEG: 1980 (mean = 68.8 mrem, median = 4.0 mrem, SE = 149.7)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
50.0	35.9	(0, 443)	30.2	2.99E+04	16.5	172.9
90.0	36.8	(0, 446)	30.6	3.01E+04	30.5	173.7
95.0	38.9	(0, 453)	30.8	3.11E+04	7.6	176.4
99.0	39.6	(0, 454)	33.0	3.11E+04	20.2	176.3
NS1 + NS3						
50.0	222.0	(0, 1155)	62.6	1.94E+05	300.5	441.0
90.0	317.8	(0, 1345)	118.6	2.38E+05	1542.6	489.3
95.0	326.8	(0, 1374)	130.8	2.46E+05	3059.2	499.6
99.0	363.1	(0, 1425)	162.8	2.55E+05	1834.8	506.9
NS1+ NS2 +NS3						
50.0	71.7	(0, 646)	32.8	6.64E+04	3.1	257.7
90.0	78.2	(0, 674)	35.1	7.23E+04	14.1	268.9
95.0	81.4	(0, 687)	37.2	7.51E+04	35.3	274.2
99.0	81.1	(0, 680)	37.2	7.31E+04	44.2	270.5
NS1 only						
50.0	-1.2	(0, 293)	1.8	1.30E+04	251.6	115.1
90.0	-52.9	(2, 29)	13.4	4.95E+01	0.6	7.1
95.0	E	E	E	E	E	E
99.0	E	E	E	E	E	E

**E = Too few measurements to successfully run the multiple imputation procedure*

Table 5.31. Performance of MI by percent missing from NS2, pipefitting SEG: 1980 (mean = 103.0 mrem, median = 34.5 mrem, SE = 169.1)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2+ NS1						
50.0	-10.5	(0, 423)	-11.9	2.85E+04	45.7	168.9
90.0	18.1	(0, 482)	-12.5	3.36E+04	346.6	184.3
95.0	-9.9	(0, 434)	-24.9	2.81E+04	1823.4	174.0
99.0	6.4	(0, 494)	-26.5	3.72E+04	1201.6	196.5
NS2 + NS3						
50.0	42.1	(0, 697)	-1.7	7.93E+04	24.3	281.7
90.0	136.9	(0, 1000)	37.9	1.50E+05	473.7	388.0
95.0	170.9	(0, 1139)	32.3	1.92E+05	2882.6	441.6
99.0	231.9	(0, 1228)	83.3	1.98E+05	8295.9	455.9
NS2 + NS1 + NS3						
50.0	36.6	(0, 658)	-2.3	7.01E+04	103.7	264.9
90.0	50.3	(0, 741)	-22.8	8.97E+04	409.3	300.3
95.0	65.9	(0, 809)	-23.2	1.05E+05	1125.7	326.7
99.0	65.9	(0, 847)	-29.8	1.19E+05	665.1	346.2
NS2 only						
50.0	1.2	(0, 461)	0.2	3.31E+04	106.8	182.4
90.0	52.9	(0, 504)	54.9	3.08E+04	629.0	177.7
95.0	24.9	(0, 324)	103.1	9.82E+03	156.5	100.1
99.0	38.6	(0, 144)	107.9	2.24E+00	0.0	1.5

Table 5.32. Performance of MI by percent missing from NS3, pipefitting SEG: 1980 (mean = 391.7 mrem, median = 136.0 mrem, SE = 497.0)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
50.0	-209.9	(0, 825)	-119.3	1.11E+05	246.8	333.9
90.0	-296.8	(0, 558)	-132.6	5.53E+04	450.1	236.3
95.0	-300.1	(0, 462)	-130.5	3.49E+04	840.6	189.4
99.0	-315.4	(0, 390)	-130.5	2.52E+04	440.8	160.3
NS3 + NS2						
50.0	-268.9	(0, 549)	-100.0	4.74E+04	18.2	217.8
90.0	-278.6	(0, 480)	-99.4	3.51E+04	35.7	187.6
95.0	-286.3	(0, 444)	-99.8	2.98E+04	18.8	172.7
99.0	-287.1	(0, 439)	-100.6	2.91E+04	9.3	170.7
NS3 + NS1 + NS2						
50.0	-271.5	(0, 547)	-103.2	4.75E+04	18.8	218.0
90.0	-284.5	(0, 468)	-103.4	3.40E+04	13.5	184.4
95.0	-287.8	(0, 446)	-103.9	3.05E+04	20.1	174.6
99.0	-288.8	(0, 442)	-104.9	3.01E+04	12.3	173.4
NS3 only						
50.0	-271.5	(0, 547)	-103.2	4.75E+04	18.8	218.0
90.0	E*	E	E	E	E	E
95.0	E	E	E	E	E	E
99.0	E	E	E	E	E	E

**E = Too few measurements to successfully run the multiple imputation procedure*

Table 5.33. Performance of MI by percent missing from NS1, pipefitting SEG: 1990 (mean = 43.1 mrem, median = 3.0 mrem, SE = 104.9)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS1 + NS2						
50.0	0.9	(0, 220)	7.6	8.13E+03	1.6	90.2
90.0	11.0	(0, 287)	13.4	1.41E+04	33.0	119.0
95.0	18.5	(0, 350)	15.8	2.16E+04	66.1	147.4
99.0	5.0	(0, 231)	13.0	8.74E+03	2.5	93.5
NS1 + NS3						
50.0	-7.7	(0, 190)	0.0	6.26E+03	1.6	79.1
90.0	-7.7	(0, 205)	0.8	7.53E+03	36.4	87.0
95.0	-11.7	(0, 173)	0.4	5.26E+03	19.9	72.7
99.0	-18.9	(0, 152)	-1.0	4.29E+03	17.9	65.7
NS1 + NS2 + NS3						
50.0	-1.0	(0, 210)	5.0	7.40E+03	2.0	86.1
90.0	4.2	(0, 251)	7.3	1.08E+04	2.0	104.1
95.0	4.3	(0, 257)	7.0	1.15E+04	14.8	107.1
99.0	-0.1	(0, 216)	6.7	7.80E+03	7.2	88.4
NS1 only						
50.0	-4.9	(0, 204)	-0.2	7.20E+03	2.8	84.9
90.0	40.3	(0, 474)	3.3	3.97E+04	71.3	199.5
95.0	14.9	(0, 438)	-2.6	3.65E+04	988.2	194.2
99.0	-10.7	(0, 309)	-3.0	1.89E+04	882.3	141.3

Table 5.34. Performance of MI by percent missing from NS2, pipefitting SEG: 1990 (mean = 53.2 mrem, median = 22.0 mrem, SE = 97.3)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS2 + NS1						
50.0	-8.6	(0, 244)	-15.4	1.04E+04	0.6	102.1
90.0	-13.6	(0, 228)	-18.6	9.32E+03	2.1	96.5
95.0	-10.6	(0, 251)	-19.0	1.13E+04	18.0	106.4
99.0	-12.1	(0, 238)	-19.8	1.01E+04	0.3	100.4
NS2 + NS3						
50.0	-12.4	(0, 207)	-13.4	7.24E+03	0.7	85.1
90.0	-22.9	(0, 158)	-17.8	4.28E+03	7.1	65.5
95.0	-25.0	(0, 150)	-18.8	3.91E+03	7.7	62.6
99.0	-25.1	(0, 150)	-19.8	3.89E+03	2.6	62.4
NS2 + NS1 + NS3						
50.0	-12.4	(0, 224)	-16.6	8.76E+03	0.7	93.6
90.0	-15.7	(0, 211)	-19.0	7.85E+03	1.1	88.6
95.0	-15.0	(0, 210)	-18.5	7.75E+03	1.7	88.1
99.0	-16.1	(0, 208)	-18.8	7.67E+03	1.1	87.6
NS2 only						
50.0	-0.3	(0, 236)	21.2	8.65E+03	103.3	93.7
90.0	-1.6	(0, 263)	16.0	1.16E+04	110.3	108.1
95.0	-5.2	(0, 180)	13.8	4.36E+03	157.9	67.4
99.0	81.2	(0, 308)	115.8	7.83E+03	17.1	88.6

Table 5.35. Performance of MI by percent missing from NS3, pipefitting SEG: 1990 (mean = 31.0 mrem, median = 3.0 mrem, SE = 66.7)

% Missing from NS1	Bias of Mean (mrem)	95% CI of Mean (mrem)	Bias of Median (mrem)	Within Variance	Between Variance	Standard Error (mrem)
NS3 + NS1						
50.0	10.4	(0, 247)	0.0	1.10E+04	1.8	105.0
90.0	15.3	(0, 296)	-1.0	1.63E+04	31.6	127.8
95.0	12.4	(0, 267)	-0.6	1.30E+04	9.4	114.2
99.0	12.7	(0, 278)	-1.0	1.43E+04	58.3	120.0
NS3 + NS1						
50.0	15.6	(0, 224)	9.5	8.26E+03	1.4	90.9
90.0	21.6	(0, 241)	18.2	9.27E+03	12.1	96.3
95.0	20.7	(0, 239)	18.8	9.18E+03	7.7	95.9
99.0	21.0	(0, 238)	18.2	9.01E+03	11.8	95.0
NS3 + NS1 + NS2						
50.0	12.6	(0, 233)	4.2	9.35E+03	0.2	96.7
90.0	16.6	(0, 284)	4.2	1.45E+04	33.0	120.6
95.0	15.6	(0, 258)	5.2	1.17E+04	13.4	108.3
99.0	14.9	(0, 255)	4.8	1.15E+04	5.1	107.1
NS3 only						
50.0	-2.5	(0, 153)	0.0	4.03E+03	3.9	63.5
90.0	6.8	(0, 186)	0.0	5.68E+03	43.2	75.7
95.0	-8.8	(0, 81)	12.0	9.07E+02	2.7	30.2
99.0	-27.9	(0, 7)	1.2	4.00E+00	0.0	2.0

5.4. Discussion

5.4.1. Characterizing the Population-level Ten-Year Exposure Profile

In this section, the limited exposure measurements available over a ten-year period for the study population at the shipyard of interest are supplemented with measurement data collected at separate shipyards. This exercise reflected a scenario in which exposure data for the study population of interest are limited but more complete data exists for a similar, but separate, facility. The desire to combine data from multiple locations to better characterize exposure must be balanced against the possibility for exposure misclassification if the exposure patterns are significantly varied between sites. The exposure variability that exists

within and between facilities can mean that combining data across sites can worsen the accuracy of the exposure estimates. Thus, prior to performing any analyses that requires a combination of data, an attempt should be made to understand the exposure patterns within each facility and how they compare across sites.

The facilities used in this exercise are all naval shipyards. As shown in several of the abovementioned tables, many of the jobs that were observed to have the greatest number of collected measurements or the greatest number of employed workers were similar across the yards. The U.S. Navy was known for having strict standards when it came to the materials workers used and the way work tasks were performed. Thus, it might be expected that exposure levels, for a given time period, would be similar across the yards. However, although exposure levels for all three yards were shown to decrease over time, the variability in exposure levels between yards changed with time as well. In the 1980s, the exposure levels between the three shipyards were more varied than was observed in the 1990s. The exposure levels were also higher and contained fewer 0 mrem values in the 1980s.

Thus, two sources of exposure variability exist that need to be considered: variability between yards during the same time period, and variability between time periods. The results showed that while the exposure estimates for both time periods performed reasonably well, the greater biases were observed when working with data from the 1980s. This suggests that the performance of an MI method for estimating missing exposure data may be influenced by the time period of interest. If the exposure data for the particular timeframe under study are less varied between sites, the estimates will likely be more accurate. While this comparison is

something that can often be known in a real-world scenario, this exercise emphasizes the importance of gathering as much information on the exposure patterns of each potential facility as possible.

There were also significant differences observed in the work population across the yards, particularly worker race and education level. These observations highlight the importance of understanding how the different facilities might vary from one another with regards to the worker population and how those variations might limit the potential analyses. For instance, in this exercise, the race variable was removed from the MI models because it was too varied between yards. Since only NS1 had a significant population of black workers, keeping race in the model would have limited the use of the data from NS2 and NS3. In doing so, however, the assumption that has to be made that there were no differences in exposure by race. While this is not likely to be completely true, race was shown to have less of an effect on exposure level as compared to other variables utilized in the MI models.

Differences in both exposure patterns and work population should therefore be expected to exist between yards, both within a given time period and over time. Ultimately these differences may prove to be less of a drawback than working with only very limited data, but the potential for misclassification is still there. The results generally indicated that using data from all three shipyards produced exposure estimates that fell in-between the estimates produced from using only two of the three yards. While this often meant that these estimates were more biased than one combination of two yards, it also meant that the estimates were less biased than the other combination of two yards. In a real-world scenario, it may not be

clear which combination of two shipyards will produce the less biased estimates. Thus, using the data from all three shipyards (or, more generally, as much data as are available) is likely the best method. Including all available data will often reduce some of the most extreme bias produced by the facility that is least similar to the one of interest. Very rarely did the estimates that were produced by using data from all three shipyards end up being the most biased estimates. This was true for estimates of the mean, median, and variance. Unless it is extremely obvious that one facility is much more similar to the one under investigation, using data from multiple facilities is preferred.

5.4.2. Characterizing the Ten-Year Exposure Profile of an SEG

In this section, the limited exposure measurements available over a ten-year period for the pipefitting SEG at the shipyard of interest are supplemented with measurement data collected at separate shipyards. The same general considerations that were discussed in the previous sections remain true here. In addition, the study population of interest has now become smaller and more specific, which may affect the amount of variability observed within and between yards.

Comparably to the prior section, the exposure levels between yards were shown to become more similar over time. Once again, exposures in the 1980-1990 time period varied more between yards as compared to the 1990-2000 time period. Exposure levels and the number of workers employed generally decreased over time for each yard and the number of measurements collected generally increased over time; however, these numbers fluctuated within and across yards over the twenty total years. These fluctuations may be responsible

for some of the observed differences in exposure levels between yards. These patterns are examples of the type of information that would be helpful to have prior to beginning any analyses that require combining data.

Also like the previous section, significant differences were observed in the work population across the yards, and again, particularly worker race. Once again, the race variable was removed from the MI models. Differences in the exposure data were also observed within and between years. Exposure levels for the 1990-2000 timeframe were lower and contained a higher percentage of zero values. There was also less variability among the data for this more recent timeframe.

Combing data from all three shipyards again produced estimates that generally fell within the estimates produced from using only two of the yards. This further suggests that using all available data should be a preferred approach unless additional information suggesting otherwise exists. Although exceptions were observed, the estimates of the mean and median for the pipefitting SEG for a given shipyard and timeframe tended to be less biased than the estimates for the entire study population for the same shipyard and timeframe; the confidence intervals were also generally narrower. This may suggest, as might be expected, that the variability within and between an SEG at different facilities is less than the variability of the overall work population within and between the same facilities.

5.4.3. Characterizing the One-Year Exposure Profile of an SEG

In this section, the limited exposure measurements available over a one-year period for the pipefitting SEG at the shipyard of interest are supplemented with measurement data collected at separate shipyards. The study population of interest has now become even smaller and more specific by focusing on just one year at a time, which may again affect the amount of variability observed within and between yards.

In contrast to the prior section, the time period of interest in this section was one year: either 1980 or 1990. When looking at the ten-year periods, it was noted that the exposure levels of the SEG were more varied over 1980-1990; however, much of that variability was due to the first half of the decade. In the later years of that decade, the exposure levels became more similar so that the overall variability was a reflection of both extremes. However, when focusing on only one year, the variability within and between shipyards (either large or small) becomes more pronounced.

In the year 1980, the exposure levels between the yards appeared to differ. There was also a large difference between the mean exposure level by quarter, with some quarters left empty. The variability within each yard was also quite high and the numbers of available measurements were rather low. For two of the three shipyards, 95% and 99% missing data meant that only one or two measurements were available for imputation; this meant that when using only the available measurements from that shipyard, the percentage of missing data could be set no higher than 90% missing.

In the year 1990, the exposure levels between yards have become more similar and the variability has decreased. The number of measurements collected also became much larger. When looking at the work population, in 1990 the workers were generally older and more educated. The distribution of race looked similar to how it did in 1980 with the exception of the missing race information for NS3 in 1980. Because of this, race was again not included as a variable in the MI models.

These two years represent different time periods in respect to similarities between yards. The estimates of the mean, median, and variance were all generally less biased when using combined data from the year 1990; the confidence intervals were also narrower. Because the exposure levels varied so greatly between yards in 1980, the estimated exposure levels from the MI models were often greatly biased. Even the estimates obtained from using exposure data from all three yards were largely inaccurate. Thus, in addition to the recommendation of using all available data when combining exposure measurements, it would also be best to estimate exposures for as long a period of time as possible. In this exercise, estimating exposures over one year resulted in the use of exposure data that were extremely varied by facility; when a ten-year time period was used instead, the variability in exposures between yards was attenuated.

5.5. Conclusion

The analyses in Chapter 5 emphasized the importance of considering not only exposure patterns within a given facility but also across facilities. It examined the true generalizability of exposures from one yard to the next. Differences in exposure levels, worker demographics, and job titles were observed between yards, which may limit the appropriateness of comparing across facilities. The exposure data from the outside facility (or facilities) were shown to heavily influence the estimates created for the yard of interest. Yet, sometimes this approach is necessary; in such a case, the results suggest using data from as many facilities as available, and over as long a time period as possible, to mitigate the influence of any one yard. These observations were made when using multiple imputation to estimate both population-level and SEG-level exposures.

CHAPTER 6: SUMMARY OF FINDINGS

6.1. Overview

Exposure misclassification likely exists in nearly every epidemiology study. The presence and level of misclassification is tied to the objective of the study; thus the severity of its effects will vary from one analysis to the next (Blair et al. 2007). The potential impacts on risk estimates and exposure-outcome relationships illustrate the importance of reducing misclassification whenever possible.

Missing data are also ubiquitous in epidemiology studies (Greenland & Finkle, 1995). Missing exposure data not only influence the understanding of the exposure-outcome relationship of interest but also can have an impact on the ultimate decisions and policies set forth from that understanding. In occupational exposure studies, missing exposure data can unintentionally affect worker protection policies and potentially even guide requirements for how a work task must be performed. Missing data in occupational cohorts are particularly concerning because they are often related to the generally biased sampling plans of industrial hygienists and/or the availability of historical data. Factors including the time period of interest, restrictions on the number of exposure samples that can be collected, and the anticipated exposure levels of various job titles can all influence which data are ultimately available.

To address missing data in a study, a variety of techniques are available. Multiple imputation has many advantages including the use of all available data and maintaining the true variability and uncertainty in the dataset. Given these advantages, it appears a logical

candidate for addressing missing exposure data in an occupational study. However, the performance of a multiple imputation approach under the common missing data patterns observed in occupational cohorts should be well understood prior to its widespread use. This requires first the identification of such common missing data patterns followed by the ability to examine a multiple imputation method under test conditions. Thus, for this dissertation, a large and complete dataset of radiation exposures of naval shipyard workers was used to investigate the applicability of multiple imputation for addressing missing data in an occupational cohort.

The research aims of this dissertation were: 1) to understand and characterize common missing exposure data patterns in occupational cohorts; 2) to test the performance of a multiple imputation approach in characterizing exposures under predetermined missing data scenarios and; 3) to comment on the observed influence missing data patterns have on the ability to accurately estimate exposures of a work population. Common missing data patterns explored in this dissertation included randomly missing, missing based on the year of sample collection, missing based on the job title of the worker and/or the expected exposure levels of various job titles, missing in order to achieve a desired percentage of sampled workers, missing based on the value of model covariates that may be potentially be related to the exposure levels, and missing based on the physical location of the facility. The performance of the multiple imputation method was evaluated by comparing the estimated exposure mean and median, the 95% confidence interval of the estimated mean, and the imputation variance to the true values of these parameters.

Ultimately, examining the relative changes in performance between trials, rather than the absolute values of the measures of accuracy, gave a better indication of how missing data patterns might influence exposure estimates and impact the abilities of the multiple imputation method. In general, the multiple imputation approach performed well in estimating exposure levels for the population of interest. This held true even when the percentage of missing data was very high (up to 99% missing). A few general patterns emerged from these analyses, which tie together nicely the second aim of testing the performance of multiple imputation and the third aim of commenting on the observed influence of missing data patterns. Changes in exposure levels over time should be well understood, particularly if there is a need to use exposure data from a different timeframe than the one of interest. A sampling plan in which the highest exposed workers are oversampled may not affect the accuracy of the exposure estimates as much as may be believed. Collecting multiple samples on the same worker can result in less biased estimates of the SEG exposure levels; however, the number of samples necessary to accurately characterize the exposure profile of a given SEG will vary based on the homogeneity of the workers within that exposure group. Giving consideration to additional model covariates that may be related to exposure can result in improved homogeneity of exposures within assigned SEGs. Finally, when estimating exposures for a specific facility using data combined from multiple locations, it is best to use data from as many different facilities as available and to estimate exposures over as long a time period as possible.

6.1.1. Estimating Population-level Exposures

The specific aims of Chapter 3 were to understand common missing data patterns in large occupational cohorts, examine the effect these patterns have on the ability to accurately characterize population-level exposures, and to test the performance of a multiple imputation approach in estimating such population exposures using a real occupational dataset with significant amounts of artificially missing exposure data. To explore these aims, three groups of analyses were carried out; each group differed in how the missing data were generated. In the first section, data were selected randomly from the overall study population at increasing percentages of missingness in order to test how well multiple imputation performed as the percentage of missing data grew larger. It is not uncommon in occupational cohorts for exposures to be characterized using samples collected on a small percentage of the population. The results of the analyses in this section suggest that an MI approach can perform well when data are missing randomly, even when the percentage of missing data is high (up to 99% missing). Thus, population-level exposures may be reasonably estimated when only a fraction of the work population has been randomly sampled. These analyses also confirmed a major advantage of multiple imputation – that is, that the total imputation variance is similar to the true variance of the data.

In the second section, data were selected to be missing based on the collection date of the sample. In many scenarios, the availability of exposure data declines as the time period of interest grows earlier. In addition, exposure levels, particularly in occupational settings, tend to decrease over time. This can result in the need to use more recent, lower exposure data to help characterize earlier, higher exposure levels. An important conclusion from this analysis

is that it is necessary to explore the potential known differences between those workers with available measurement data and those without. In this section, workers who were employed during an earlier time period were more likely to have missing data; these workers were also more likely to have higher exposures, based on observed trends in occupational exposures. Acknowledging such patterns allowed for more accurate estimates of the population exposure levels (for example, by using the HML analysis). Differences between populations with missing data and those without are not uncommon in exposure studies but may not always be considered during analysis. Information on such differences can potentially be found in the published literature, in previous measurements, or in data from a similar facility or industry.

In the third and final section, exposure data were selected to be missing based on the job titles of the worker population. When working with data collected by an industrial hygienist, the missing data patterns may be based on the perceived exposure levels of each job title. The information the hygienist uses to design the sampling plan (prior exposures, peer-reviewed literature, etc.) may vary, but the ultimate concern remains the same: is the available data biased in some way that will be reflected in the estimated exposures. Several different sampling strategies were simulated in this section in an attempt to capture some of these potential biases. Despite the differences in sampling plans between analyses, all performed reasonably well. This suggests that there are a number of appropriate sampling designs and that even when available data are biased by job title, multiple imputation is a viable option.

6.1.2. Characterizing the Exposure Profile of an SEG

The specific aims of Chapter 4 were to investigate the performance of similar exposure groups under various conditions, examine how SEGs are affected by changes in the sample size, explore additional workplace variables that may influence the homogeneity of an SEG, and test the performance of a multiple imputation approach in estimating SEG-level exposures. The analyses in this section allowed for a better understanding of the factors that influence the homogeneity of an SEG and thus influence the ability to accurately characterize the exposures of the work force.

In the first section, it was acknowledged that as the percentage of missing data increases, the number of available samples per year within an SEG grows smaller, sometimes to only one or two available samples. This can make analyses using a modeling approach difficult. One solution is to create broad time intervals (such as the 5- and 10-year bins used in this section) to use in the model. Grouping data together into fewer bins increases the sample size per bin. Although the performance of all three models did drop as the percentage of missing data increased, the model using the broadest time intervals produced the least biased estimates of exposure. However, this strategy comes at a cost, as there was an observed underestimation of the total variance. Thus, another important overall conclusion is that when deciding which approach to take in addressing missing data, the ultimate goals of the study should be considered; the relative importance of factors such as unbiased exposure estimates versus more a more accurate estimate of variance should be weighed.

In the second section, the total number of samples collected within an SEG was varied by the number of workers sampled and the number of samples collected per worker, in an attempt to

capture a variety of plausible sampling plans. Decisions over sampling plans are ones industrial hygienists and researchers alike are commonly faced with, as there are often a set number of total samples that can be feasibly collected. The goal then becomes to design a sampling plan that captures enough variability to answer the intended study questions. The analyses in these sections suggest that collecting multiple samples per worker, when possible, can result in less biased estimates, particularly for variance. However, the homogeneity of exposures within one SEG may vary from another, such that the number of samples necessary to accurately characterize the exposure profile for that SEG may not be the same for the next. Because of this, it is important to understand the determinants of exposure for a particular workforce and then assign workers to SEGs based on the most influential factors.

The final section of this chapter attempted to do just that – consider additional variables that may assist in defining SEGs. A relationship between worker birth year and mean exposure level was observed for all three SEGs examined and at all three shipyards. Removing birth year from the model also resulted in the lowest estimates of variance, suggesting that some of the within-SEG variability is due to this variable. The hypothesis is that birth year is serving as a surrogate for additional workplace exposure information that is not available, such as work task within a job title. This exercise illustrates the potential for seemingly unrelated, but perhaps easily available, variables to provide information regarding the exposure levels of the work population. Given the heterogeneity of exposure by birth year within a supposed SEG, it is also suggested that solely using job titles to define SEGs, unless the absolute only option, should be avoided. This is currently a common practice, although there has been

much discussion on the topic (Weinkam et al. 1991). Indeed, homogeneity of exposures within an SEG will likely continue to be explored for some time.

6.1.3. Developing Exposure Estimates Using Surrogate Data

In Chapter 5, the specific aims were to compare between shipyards the exposure profile of naval shipyard workers during various time periods and to test the performance of a multiple imputation approach in estimating exposure levels for the shipyard of interest when exposure data from multiple shipyards are combined.

In the first section, the population-level exposures over a ten-year period for one shipyard were estimated using data combined from at least two, and in some cases all three, yards. This exercise represented yet another plausible scenario, a situation in which exposure data for the site of interest are limited and additional data are available at similar but separate facilities. The analyses in this section emphasized the importance of considering not only exposure patterns within a given facility but also across facilities. It examined the true generalizability of exposures from one yard to the next. Differences in exposure levels, worker demographics, and job titles were observed between yards, which may limit the appropriateness of comparing across facilities. The exposure data from the outside facility (or facilities) were shown to heavily influence the estimates created for the yard of interest. Yet, sometimes this approach is necessary; in such a case, the results of this section suggest using data from as many facilities as available to mitigate the influence of any one yard.

The remaining two sections of this chapter examined the potential of using data from multiple yards to characterize the exposure profile of a specific SEG. As compared to the population-level analyses, working within an SEG results in a smaller, more specific population, resulting is less overall variability. However, as demonstrated in Chapter 4, heterogeneity still exists even with an exposure group. Once again it was necessary to consider not only exposure patterns within a yard but across yards. Like for the population-level analyses, it was also important to consider the time period over which the data are to be combined. As demonstrated in the analyses within this chapter, exposures during some time periods (here, the 1990-2000 decade; the single year 1990) are more similar between yards than during other years (1980-1990 decade; the single year 1980). In addition to using all available data when combining exposure measurements across yards, the results of this chapter also suggest that estimating exposures over as long a time period as possible will produce more accurate results. For instance, estimating exposures over a one-year period resulted in the use of exposure data that were extremely varied by facility; this variability was attenuated when a 10-year time period was used instead.

6.2. Strengths and Limitations

6.2.1. Strengths

The dataset of radiation exposures of naval shipyard workers used for the analyses in this dissertation included over one million daily radiation measurements (and 100,000 annual radiation measurements) collected on approximately 13,8000 workers from three shipyards over thirty years. The size and completeness of these data allows for the ability to carry out analyses that would not be possible with a smaller dataset. By working with a large dataset

that contains no missing exposure data, many plausible missing data patterns observed in occupational cohorts could be simulated. This included more nuanced examinations such as: trends in exposure levels over both short and long periods of time; the effects of creating subsets of exposure data for various analyses; the comparison of observed results across several SEGs, all of which had no missing data; and the effect of combining exposure data from multiple yards.

The potential for using a multiple imputation approach to estimate missing occupational exposure data could also assessed by working with this dataset. The true values of the desired exposure metrics were easily obtained, allowing for an accurate evaluation of the estimates generated by the multiple imputation approach.

Finally, given the nature of the monitoring strategy employed by the Navy, this dataset represents an unbiased sampling design. Workers were instructed to wear a radiation dosimeter whenever they entered a potential radiation exposure area, regardless of the actual resulting exposure level; a badge was worn every time the employee re-entered such an area. Thus, this dataset not only captures all shipyard employees who were ever potentially exposed to radiation during their work on the submarines but also every single occasion for exposure each individual worker had.

6.2.2. Limitations

The study population and dataset used for these analyses have some unique qualities that may make them less generalizable to other occupational cohorts than desired. Shipyard workers,

similar to those in the construction industry, have a distinctive work pattern that involves migrating around a facility (or ship) to various workstations and performing their work tasks in moments of various lengths of time. Their daily schedule is based around the required work to be completed and may vary significantly from day to day, even from hour to hour. This is in contrast to the more conventional work populations of industry – factory workers who perform the same tasks, in the same general location, all day, each day. Shipyard workers' exposures may therefore be more sporadic and variable, which may require altered sampling and analysis strategies. However, this fact should actually strengthen the argument for multiple imputation. If the method performs well even when the work population is so varied, it will likely improve in performance in a scenario in which work tasks are more homogenized.

The dataset is also larger and more complete than most occupational exposure dataset would be. While this made the cohort an excellent candidate for this dissertation project (thus also making it a strength of the study), it also makes it more difficult to relate the available sample sizes, especially at high percentages of missing data, to what would be feasibly observed in a smaller study.

Multiple imputation has its own set of limitations, which were discussed in Chapter 2. It is a more computationally advanced method for addressing missing data as compared to some of the more common approaches like complete-case analysis. While the assumed missing data mechanism when working with multiple imputation (MAR) is less restrictive than the MCAR assumption needed for a deletion method, it still requires an assumption that the reasons the

data are missing are unrelated to the values of the missing data themselves. This might not be a plausible assumption in some cases and, in all cases, requires some careful consideration of the data on the part of the researcher.

6.3. Public Health Implications

As stated previously, the overall goal of this dissertation project was to investigate a method for developing improved exposure estimates in an effort to reduce the potential for exposure misclassification. The effects of misclassification on risk estimates and exposure-outcome associations have been discussed. This is a well acknowledged but difficult to address problem that affects all epidemiology studies regardless of the study population. Working with an occupational cohort brings an additional set of unique challenges to the exposure assessment. In order to properly estimate exposures for this population, and reduce misclassification, these challenges need to be recognized and addressed.

The first aim of this project was to understand and characterize common missing exposure data patterns in occupational cohorts. By identifying and evaluating the missing data patterns discussed in this dissertation, which are believed to be some of the most commonly faced when working with occupational exposure data, there is an opportunity to better understand and anticipate the impacts such patterns will have in future occupational exposure studies. Missing data patterns are often present but not detected; this project illustrates how learning to seek out patterns in the data prior to conducting an analysis can significantly improve exposure estimates.

This project also supports the use of multiple imputation as an approach for addressing missing occupational exposure data. As noted in Chapter 1, there are a number of potential techniques when working with missing data and all have their strengths and weaknesses. Multiple imputation is gaining popularity as an approach with many advantages and its practicality when working with occupational exposure data should be explored. By testing the performance of a multiple imputation approach under various plausible missing data patterns, this dissertation has demonstrated that multiple imputation is an appropriate choice for working with occupational exposure datasets, particularly when the missing data patterns have been well characterized.

Finally, this research highlights the potential disconnects between the original purposes of industrial hygiene exposure data in compliance determinations and their possible eventual use in occupational epidemiology studies. Since the ultimate objectives of these two analyses differ significantly, the manner in which the exposure data are used and interpreted can also vary. Rather than unknowingly work with a biased dataset, researchers should be encouraged to understand the original intent of the available exposure data and how the missing data patterns might affect their own results. When designing future sampling plans, both industrial hygienists in the field and researchers should consider the potential future uses of the dataset. A deeper understanding of the fundamental disciplines of industrial hygiene, epidemiology, and biostatistics will allow for all investigators working with occupational exposure data to collect and interpret data more efficiently.

6.4. Conclusion

This dissertation examined some of the common missing data patterns that emerge in occupational cohorts, explored how these patterns can influence the ability to accurately estimate exposure, and supported the use of a multiple imputation approach for addressing missing exposure data. In order to properly characterize the exposure profile of a population of interest, the patterns of missing data need to be identified and the potential differences between available and missing data should be explored. When selecting an approach for addressing missing data, the overall objectives of the analyses should be considered. Multiple imputation is one approach that offers many advantages and has been demonstrated to work well with occupational exposure data. Ultimately, the ability to accurately estimate exposures for a population will depend on the completeness of the data and the level to which the missing data are characterized. A solid understanding of the fundamentals of industrial hygiene, epidemiology, and biostatistics will assist in interpreting and developing exposure datasets.

CHAPTER 7: REFERENCES

Armstrong B. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med.* 1998;55:651-656.

Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med*. 2007;50(3):199-207.

Burdorf A, Van Tongeren M. Commentary: Variability in workplace exposures and the design of efficient measurement and control strategies. *Ann Occup Hyg*. 2003;47(2):95-99.

Cember H, Johnson TE. (2009). *Introduction to Health Physics 4th Ed.* New York. McGraw Hill Medical.

Checkoway H, Dement JM, Fowler DP, Harris RL, Lamm SH, Smith TJ. Industrial hygiene involvement in occupational epidemiology. *Am Ind Hyg Assoc J.* 1987;48(6):515-523.

Checkoway H, Pearce N, Kriebel D. (2004). *Research Methods in Occupational Epidemiology 2nd Ed.* Oxford. Oxford University Press, Inc.

Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488-495.

Corn M, Esmen NA. Workplace exposure zones for classification of employee exposures to physical and chemical agents. *Am Ind Hyg Assoc J*. 1979;40(1):47-57.

Correa-Villasenor, A. "A case-control study of mesothelioma in the shipyard industry." Diss. The Johns Hopkins University, 1987.

Daniels RD, Taulbee TD, Chen P. Radiation exposure assessment for Portsmouth naval shipyard health studies. *Radiat Prot Dosimetry*. 2004;111(2):139-150.

Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat Med*. 2003;22(4):545-557.

Dinardi SR, ed. (2003). *The occupational environment: Its evaluation and control and management 2nd Ed*. Fairfax. AIHA Press

Enders CK. Analyzing longitudinal data with missing values. *Rehabil Psychol*. 2011;56(4):267-288.

Federal Register 57(104). Guidelines for exposure assessment. May 29 1992:22888-22938.

Fielding S, Fayers P, Ramsay C. Predicting missing quality of life data that were later recovered: An empirical comparison of approaches. *Clin Trials*. 2010;7(4):333-342.

Gollnick DA. (2006). *Basic radiation protection technology 5th ed.* Altadena. Pacific Radiation Corp.

Graham, JW. Missing data analysis: Making it work in the real world. *Annu Rev Psychol.* 2009;60:549-576.

Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol.* 1995;142(12):1255-1264.

Harris RL. Guideline for collection of industrial hygiene exposure assessment data for epidemiologic use. *Appl Occup Environ Hyg*. 1995;10(4):311-316.

Hawkins NC, Evans JS. Subjective estimation of toluene exposures: A calibration study of industrial hygienists. *Appl Ind Hyg*. 1989;4(3):61-68.

Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: Expectations vs observations. *Int J Epidemiol*. 2005;34(3):680-687.

Jurek AM, Greenland S, Maldonado G. How far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?. *Int. J Epidemiol.* 2008;37:382-385.

Kromhout H, Symanski E, Rappaport SM. A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Ann Occup Hyg*. 1993;37(3):253-270.

Last J. (2001). *A dictionary of epidemiology 4th ed*. New York. Oxford University Press.

Loomis D, Kromhout H. Exposure variability: Concepts and applications in occupational epidemiology. *Am J Ind Med*. 2004;45(1):113-122.

Ma J, Raina P, Beyene J, Thabane L. Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study. *Open Access Med Stat.* 2012;2:93-103.

Matanoski GM, Tonascia JA, Correa-Villasenor A, Yates KC, Fink N, Elliott E, Sanders B, Lantry D. Cancer risks and low-level radiation in U.S. shipyard workers. *J Radiat Res*. 2008;49(1):83-91.

Mishra S, Khare D. On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. *J Med Stat Inform*. 2014;2(9):1-7.

Nieuwenhuijsen MJ. "Introduction to exposure assessment." *Exposure assessment in occupational and environmental epidemiology*. Ed. Mark Nieuwenhuijsen. London, 2003. 3-19. Print.

NRC (National Research Council). (1983). *Risk Assessment in the Federal Government: Managing the process.* Washington D.C. National Academy Press.

NRC (National Research Council). (2009). *Science and Decisions: Advancing Risk Assessment*. Washington D.C. National Academy Press.

Perkins, JL. (1997). *Modern Industrial Hygiene Recognition and Evaluation of Chemical Agents Vol. I.* New York. Van Nostrand Reinhold.

Pigott, TD. A review of methods for missing data. *Educ Res Eval.* 2001;7(4):353-383.

Raghunathan T. What do we do with missing data? Some options for analysis of incomplete data. *Annu Rev Public Health*. 2004;25:99-117.

Ramachandran G, Banerjee S, Vincent JH. Expert judgment and occupational hygiene: Application to aerosol speciation in the nickel primary production industry. *Ann Occup Hyg.* 2003;47(6):461-475.

Rappaport SM. Assessment of long-term exposures to toxic substances in air. *Ann Occup Hyg*. 1991;35(1):61-121.

Rappaport SM, Kromhout H, Symanski E. Variation of exposure between workers in homogeneous exposure groups. *Am Ind Hyg Assoc J*. 1993;54(11):654-662.

Rubin, DB. Inference and missing data. *Biometrika*. 1976;63(3):581-590.

Rubin, DB. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York. J. Wiley & Sons.

Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychol Methods.* 2002; 7(2):147-177.

Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivar Behav Res*. 1998;33:545-571.

Seel EA, Zaebst DD, Hein MJ, Liu J, Nowlin SJ, Chen P. Inter-rater agreement for a retrospective exposure assessment of asbestos, chromium, nickel and welding fumes in a study of lung cancer and ionizing radiation. *Ann Occup Hyg*. 2007;51(7):601-610.

Stern FB, Waxweiler RA, Beaumont JL, Lee ST, Rinsky RA, Zumwalde RD, Halperin WE, Bierbaum PJ, Landrigan PJ, Murray Jr. WE. A case-control study of leukemia at a naval nuclear shipyard. *Am J Epidemiol*. 1986;123(6):980-992.

Sterne, JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Br Med J.* 2009;338:b2393.

Stewart P, Stenzel M. Data needs for occupational epidemiologic studies. *J Environ Monit.* 1999 Aug;1(4):75N-82N.

Stewart P, Stenzel M. Exposure assessment in the occupational setting. *Appl Occup Environ Hyg*. 2000;15(5):435-444.

Symanski E, Kupper LL, Hertz-Picciotto I, Rappaport SM. Comprehensive evaluation of long term trends in occupational exposure: part 2. Predictive models for declining exposures. *Occup Environ Med.* 1998a;55:310-316.

Symanski E, Kupper LL, Rappaport SM. Comprehensive evaluation of long term trends in occupational exposure: part 1. Description of the database. *Occup Environ Med.* 1998b;55:300-309.

Vadali M, Ramachandran G, Mulhausen J. Exposure modeling in occupational hygiene decision making. *J Occup Environ Hyg*. 2009;6(6):353-362.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton. CRC Press.

van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011: 45(3): 1-67

von Hippel, PT. How many imputations are needed? A comment on Hershberger and Fisher (2003). *Struct Equ Modeling*. 2005;12(2):334-335.

Weinkam JJ, Rosenbaum WL, Sterling TD. A practical approach to estimating the true effect of exposure despite imprecise exposure classification. *Am J Ind Med*. 1991;19(5):587-601.

Werner MA, Attfield MD. Effect of different grouping strategies in developing estimates of personal exposures: Specificity versus precision. *Appl Occup Environ Hyg*. 2000;15(1):21-25.

Zaebst DD, Seel EA, Yiin JH, Nowlin SJ, Chen P. Summary of retrospective asbestos and welding fume exposure estimates for a nuclear naval shipyard and their correlation with radiation exposure estimates. *J Occup Environ Hyg.* 2009;6(7):404-414.

Pamela Dopart

615 N. Wolfe St. E6628, Baltimore, MD 21205
(415) 999-3206 • pdopart1@jhu.edu
Born: April 24, 1984, Rahway NJ

EDUCATION

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD — *2010-2015*
Ph.D. candidate, Environmental Health Sciences — *(expected)*

University of Michigan School of Public Health, Ann Arbor, MI — *2006-2008*
M.P.H., Industrial Hygiene

James Madison University, Harrisonburg, VA — *2002-2006*
B.S., Chemistry

RESEARCH EXPERIENCE

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Dissertation Research, 2010-2015

- Research experience in exposure assessment methodologies, including within the context of occupational epidemiology studies.
- Thesis research is centered on the exploration of a multiple imputation approach for developing exposure estimates from limited available measurement data.
- Thesis research involves work with a large, longitudinal dataset of radiation exposures of naval shipyard workers.
- Relevant coursework completed in industrial hygiene, exposure assessment, epidemiology, biostatistics, and risk assessment.

University of Michigan School of Public Health, Ann Arbor, MI
Graduate Research, January – May 2008

- Assistant researcher on a cross-sectional study on fatigue, discomfort, and musculoskeletal disorders (MSDs) associated with prolonged standing and walking at a large automobile manufacturing plant.

Graduate Research, September 2006 – May 2007

- Assistant researcher on a population-based case-control study examining the relationship between arsenic levels in drinking water and cases of bladder cancer.

CERTIFICATES AND TRAINING

Certificate, Risk Sciences and Public Policy, Department of Health Policy and Management, The Johns Hopkins Bloomberg School of Public Health

Hazardous Substance Academic Training Program (HSAT), University of Michigan School of Public Health

TEACHING AND WORK EXPERIENCE

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Teaching Assistant

- First Term 2011: Fundamentals of Occupational Health
- Third Term(s) 2012-2014: Introductory Principles of Environmental Health

Cardno ChemRisk, San Francisco, CA
Associate Health Scientist, June 2008 – June 2010

Sandia National Laboratories, Albuquerque, NM
Industrial Hygiene Intern, May – August 2007

PUBLICATIONS AND ABSTRACTS

Madl, A.K., D.M. Hollins, K.D. Devlin, E.P. Donovan, **P.J. Dopart,** P.K. Scott, and A.L. Perez (2014). Airborne asbestos exposures associated with gasket and packing replacement: A simulation study and meta-analysis. *Reg Tox Pharmacol.* 69(3):304-319.

D. M. Cowan, **P. Dopart,** T. Ferracini, J. Sahmel, K. Merryman, S. Gaffney, D. J. Paustenbach (2010). A cross-sectional analysis of reported corporate environmental sustainability practices. *Reg Tox Pharmacol*. 58(3):524-538.

P.J. Dopart, P.H. Dalton, C. Maute, P.S. Lees (2013). Understanding Factors Related to Between- and Within-Subject Variation in an Effort to Better Define SEGs. Abstract 274, 23rd Conference on Epidemiology in Occupational Health Abstract Book, June 18-21, 2013, Utrecht, The Netherlands *(with oral presentation)*

P.J. Dopart, P.H. Dalton, C. Maute, P.S. Lees (2012). Exploration of Variation in SEGs in an Effort to Reduce Exposure Misclassification. Abstract WC2-02, International Society of Exposure Science 22nd Annual Meeting Abstract Book, October 28-November 1, 2012, Seattle, WA *(with oral & poster presentations)*

INVITED LECTURES

Using Multiple Imputation to Estimate Radiation Exposures. Johns Hopkins Bloomberg School of Public Health, Baltimore, MD. April 14, 2015.

Exposure, Dose-Response, and Risk Assessment. Johns Hopkins Bloomberg School of Public Health, Baltimore, MD. February 6, 2014.

Toxicity Potential of Chemical Commonly Found in Our Environment. California College of the Arts, San Francisco, CA. March 4, 2010.

Toxicity Potential of Chemical Commonly Found in Building Materials. Academy of Art University, San Francisco, CA. March 3, 2009.

PROFESSIONAL ASSOCIATIONS

American Industrial Hygiene Association (AIHA)

- Occupational and Environmental Epidemiology Committee (OEEC) member
- Student and Early Career Professionals Committee (SECP) member
- American Industrial Hygiene Foundation (AIHF) Communications Team member

International Society of Exposure Science (ISES)

HONORS AND AWARDS

Honorable Mention Award, EHS Annual Retreat Student Poster Competition, *2015*
AIHce 2014 Student Sponsorship Travel Award, *2014*
The Johns Hopkins University SPH Morgan-James Scholarship Fund, *2013*
3M Personal Safety Division's Occupational Health and Safety Scholarship Award, *2013*
University of Michigan SPH Marvin Selin Memorial Scholarship, *2008*
American Industrial Hygiene Foundation Ralph G. Smith Scholarship, *2007*
University of Michigan Environmental Health Science Student Award, *2007*
University of Michigan SPH Dean's Award, *2006-2008*
Iota Sigma Pi National Honors Society for Women in Chemistry, *inducted in 2005*

LEADERSHIP

President, Environmental Health Sciences Student Organization (EHSSO),
Johns Hopkins Bloomberg School of Public Health, 2013-2014.

President-Elect, Environmental Health Sciences Student Organization (EHSSO),
Johns Hopkins Bloomberg School of Public Health, 2012-2013.

Community Service Chair, U of M Club of Greater San Francisco, 2008-2010.

President, University of Michigan Industrial Hygiene Student Association (UMIHSA), University of Michigan School of Public Health, 2007-2008.